Effects of ChatGPT-Assisted and Claude-Assisted Instruction on the Academic Performance of Computer Science Students in Colleges of Education

Dr. JIBRIL, Mohammed; ALHASSAN, Mohammed Babatifin

9jibrilmohammed@gmail.com; alhassan.mb@kwasued.edu.ng

Department of Educational Technology, University of Ilorin, Ilorin, Nigeria
Department of Technology Education, Kwara State University of Education, Ilorin

https://orcid.org/0009-0003-3597-151X

Abstract

Large language models (LLMs) are artificial intelligence systems built on transformer architecture that use self-attention mechanisms to generate human-like text from patterns learned in massive training data; ChatGPT and Claude are among the LLMs increasingly used to support instruction across disciplines. Empirical evidence comparing these LLMs against one another, particularly in computer science education, remains limited. This study determined and compared the effects of ChatGPT-assisted and Claude-assisted instruction on the academic performance of Computer Science students in Colleges of Education in Kwara State. A quasi-experimental, pretest-posttest nonequivalent control group design was adopted, with 93, 200-level Computer Science students purposively drawn from three intact classes. Data were collected using a validated 40-item Multiple Choice Performance Test (MCPT), with a reliability coefficient of 0.76 using Kuder-Richardson (KR-20). Data were analyzed with ANCOVA at 0.05 significance level. Both AI-assisted groups significantly outperformed the conventional group (mean gains: ChatGPT = 10.71, Claude = 12.27, Conventional = 5.34), with Claude-assisted instruction yielding significantly higher posttest scores than ChatGPT-assisted instruction (26.24 to 24.71, $p = .041$). The study concludes that both ChatGPT and Claude are effective instructional tools, but Claude-assisted instruction produced significantly greater gains. The study recommends adopting Claude as a preferred supplementary instructional tool in computer science education.

Keywords: Large Language Models, ChatGPT, Claude, AI-assisted instruction, Academic performance, Computer science education

INTRODUCTION

The advancement of generative artificial intelligence has transformed instructional practices in higher education globally, with large language models (LLMs) such as ChatGPT now widely used to support teaching across many disciplines. According to a meta-analysis of experimental studies on ChatGPT use in education, factors such as the subject area, the duration of the intervention, and the way the tool is incorporated into instruction all influence the extent to which students actually benefit from it (Wu et al., 2026). In other words, the same tool can produce different results depending on how and where it is used. This dependence is especially relevant in computational disciplines, where LLMs such as ChatGPT, Claude, Gemini, and DeepSeek are increasingly used to generate code, explain concepts, and help students debug their programs, functions that make them particularly useful in courses like data structures (Li et al., 2026).

These capabilities align closely with what computer science education is trying to build in students: logical reasoning, syntax mastery, and the ability to break problems down into manageable steps. This has prompted researchers to pay growing attention to how LLMs perform in computer science classrooms specifically. Recent studies on programming and data-related courses have moved beyond simply asking

whether students like using these tools, and have started directly measuring whether they actually improve students' performance (Li et al., 2026). Similarly, a study found that ChatGPT-assisted instruction significantly improved students' academic performance, retention, and cognitive ability compared to a conventionally taught group (Ukume et al., 2026). Findings like this support a broader pattern seen across the literature: when generative AI tools are deliberately built into instruction, they can outperform traditional lecture-based teaching across a wide range of subjects and educational levels. Much of this evidence, however, centers on ChatGPT, leaving open the question of how other widely available LLMs, such as Claude, compare when applied to the same kind of instructional task.

Claude specifically has received comparatively little empirical attention alongside ChatGPT. Its developer describes its educational use as emphasizing guided reasoning, such as prompting students with questions about how they would approach a problem rather than supplying a solution outright (Anthropic, 2025), but this is a statement of design intent and not independent evidence. Independent comparisons remain limited. Claude has performed comparably to or slightly better than GPT-4 on medical-education multiple-choice questions (Mavrych et al., 2025), while a small simulated-student case study found Claude 2's tutoring responses more direct and conventional in style than ChatGPT's, even when all chatbots were prompted to teach Socratically (Santos, 2023). Early exploratory work applying Claude as an academic support tool has noted that some researchers favor it over more commonly studied LLMs precisely because of this more cautious, less answer-dispensing conversational style, reasoning that it may leave less room for students to copy unearned solutions (Divekar et al., 2024). Whether this difference in design philosophy actually produces measurable differences in student performance compared with ChatGPT-assisted instruction, however, is a question that exploratory and survey-based studies are not equipped to answer.

The limited experimental evidence directly comparing Claude with other LLMs on technical educational tasks lends some support to the idea that the two tools may behave differently in ways relevant to learning outcomes. In a study evaluating ChatGPT-4, Claude 3, and Gemini 1.5 on grading undergraduate programming assignments, Claude 3 showed a weaker correlation with instructor-assigned grades, a larger average score deviation, and more variable, less rubric-aligned feedback than the other two models (Bernik et al., 2025). Although that study examined Claude as a grading aid rather than as a direct instructional agent, its findings suggest that models can differ meaningfully in the consistency and specificity of the explanatory feedback they produce. No identified study in a developing country with limited resources has yet compared Claude against ChatGPT as competing instructional treatments in a computer science classroom, and it therefore remains unclear whether these model-level differences translate into differential gains in student academic performance.

Statement of the Problem

LLM tools are increasingly available to students and lecturers in Nigerian Colleges of Education, yet there is little empirical evidence to guide which tool should be used for instruction. Existing studies from Nigerian institutions have examined a single tool, usually ChatGPT, or have treated generative AI as one undifferentiated category, and have compared it only with conventional teaching (Bolarinwa et al., 2026; Mohammed et al., 2025; Suleiman et al., 2026; Ukume et al., 2026). No study identified has compared two named LLMs against each other, and against conventional teaching, on the academic performance of Computer Science students in Colleges of Education. Consequently, Computer Science lecturers and pre-service teachers lack empirical guidance on whether freely available LLMs such as ChatGPT and Claude differ in their effects on students' performance in specific computer science concepts. It is this gap that the present study sought to address.

Purpose of the Study

The main purpose of this study was to determine and compare the effects of ChatGPT-assisted and Claude-assisted instruction on the academic performance of Computer Science students in Colleges of Education in Kwara State. Specifically, the study:

1. Determined the effect of ChatGPT-assisted instruction on the academic performance of Computer Science students in selected computer science concepts.
2. Determined the effect of Claude-assisted instruction on the academic performance of Computer Science students in selected computer science concepts.
3. Compared the academic performance of Computer Science students taught selected computer science concepts using ChatGPT-assisted instruction and those taught using Claude-assisted instruction.

Research Questions

The following research questions guided the study:

1. What is the effect of ChatGPT-assisted instruction on the academic performance of Computer Science students in selected computer science concepts?
2. What is the effect of Claude-assisted instruction on the academic performance of Computer Science students in selected computer science concepts?
3. What is the difference between the mean academic performance scores of Computer Science students taught selected computer science concepts using ChatGPT-assisted instruction and those taught using Claude-assisted instruction?

Hypotheses

The following null hypotheses were tested at the 0.05 level of significance:

$H0_1$: There is no significant difference in the academic performance of students taught using ChatGPT-assisted instruction and those taught using the conventional method.

$H0_2$: There is no significant difference in the academic performance of students taught using Claude-assisted instruction and those taught using the conventional method.

$H0_3$: There is no significant difference in the academic performance of students taught using ChatGPT-assisted instruction and those taught using Claude-assisted instruction.

Theoretical Framework

This study is based on Vygotsky's (1978) sociocultural theory of learning. It focuses on three ideas from the theory: the zone of proximal development (ZPD), mediation and scaffolding. It is also supported by the ICAP framework of cognitive engagement (Chi & Wylie, 2014). Vygotsky (1978) argued that students learn best when a more knowledgeable person helps them with tasks they cannot yet do alone. The ZPD is the gap between what a student can do without help and what they can do with guidance. Vygotsky also noted that learning is mediated, meaning it is supported by tools, such as language or technology, that help learners understand new ideas. Wood, Bruner, and Ross (1976) called the help given within the ZPD scaffolding. Good scaffolding has two features. It is responsive, meaning the support matches what the student is struggling with at that moment. It is also gradually reduced, so that students take over the thinking themselves as they improve (van de Pol et al., 2010).

In this study, the LLM served as a mediating tool at two levels. First, the lecturer used ChatGPT or Claude to deliver instruction, generating explanations, worked examples, and step-by-step demonstrations of the four computer science concepts. Second, students were allowed to interact with the same tool directly, asking questions at the point of confusion, testing their own attempts and receiving immediate feedback. Teacher-mediated delivery provides structured support. Direct student–LLM interaction provides individualised, on-demand support that a single lecturer cannot offer to a whole class at once. In the conventional group, explanation and pacing were fixed for all students, with fewer opportunities for immediate, personalised support. Concepts such as number system conversion, algorithm and flowchart design, and programming constructs require step-by-step reasoning that benefits from this kind of responsive support.

Scaffolding is not equally effective in every form. The ICAP framework (Chi & Wylie, 2014) proposes that learning outcomes increase as students move from passive engagement (receiving information) to active (manipulating it), constructive (generating new explanations or solutions), and interactive (co-constructing understanding through dialogue). A tool that mainly supplies finished answers risks keeping students at the passive or active level. A tool that prompts students to reason through a problem, as Anthropic (2025) describes for Claude, may encourage constructive and interactive engagement. Because ChatGPT and Claude are built and fine-tuned differently, they may provide different degrees of support even when the lecturer and content are the same.

Literature Review

Large language models (LLMs) are artificial intelligence systems built on the transformer architecture, a deep learning design that uses self-attention mechanisms to process and generate human-like text by learning statistical patterns across massive volumes of training data (Mienye et al., 2025). Unlike earlier rule-based natural language systems, LLMs such as GPT (the architecture underlying ChatGPT) are pre-trained on broad, general-purpose text and can subsequently generate coherent responses without being explicitly programmed for each task (Mienye et al., 2025). This general-purpose capability is what has made LLMs, including ChatGPT (developed by OpenAI) and Claude (developed by Anthropic), attractive for instructional use: both are conversational LLMs capable of explaining concepts, answering questions, generating examples, and responding adaptively to learner prompts, even though they are built and fine-tuned differently by their respective developers and may therefore differ in how they explain concepts. It is these differences in training and fine-tuning that provide the conceptual basis for comparing two distinct LLMs, rather than a single tool, as instructional agents in the same course.

Empirical interest in how these models perform when applied to actual classroom instruction has grown. Bolarinwa et al. (2026) conducted a study among NCE III Chemistry students at the Federal College of Education (Special), Oyo, and found that ChatGPT-assisted instruction produced significant gains in students' academic performance in thermodynamics compared with conventional teaching. Not all evidence points in the same direction; Kosar et al. (2024), in a controlled experiment involving first-year undergraduate programming students, found no statistically significant difference in performance, assignment grades, or midterm exam scores between students who used ChatGPT and those who did not. Similarly, Yang et al. (2025), working with high school programming students, reported that those who used ChatGPT recorded lower flow experience, self-efficacy, and learning achievement than students taught through conventional methods. These studies suggest that the effectiveness of an LLM is neither automatic nor universal, and may depend on how the tool is implemented in instruction, a consideration directly relevant to interpreting whether ChatGPT-assisted instruction produces the expected gains in the present study.

The studies above center on ChatGPT; comparable performance-based evidence involving Claude in programming instruction specifically is far more limited. Fan et al. (2025) compared AI-assisted pair programming, human-human pair programming, and individual programming among 234 undergraduate

students in a Java web development course, using GPT-3.5 Turbo in the first year and Claude 3 Opus in the second. Relative to individual programming, the AI-assisted groups showed significantly better programming performance, matching outcomes from human-human pairing. The two models were not compared head-to-head, so the gains cannot be attributed to one tool over the other, but the study remains one of the few to link Claude specifically to a measured performance outcome rather than a grading or writing task.

Fewer studies have moved beyond a single LLM to compare multiple models against each other within the same instructional setting, but the available evidence offers useful points of comparison. Elhag et al. (2025), in a study among university students in an educational and information technology course, found that instruction supported by generative AI tools, including ChatGPT and Gemini, produced significantly higher achievement scores than conventional teaching. This finding corroborates the achievement-enhancing potential of LLM-supported instruction broadly. What remains largely unexamined, and what this body of literature collectively points toward as a logical next step, is how two specific, named LLMs, ChatGPT and Claude, compare against each other in producing academic performance gains within the same course, population, and assessment instrument.

Methodology

Design and Sample

This study adopted a quasi-experimental design, specifically the pretest-posttest nonequivalent control group design, represented as shown in the research design layout, where O_1 denotes the pretest, O_2 denotes the posttest, X_1 denotes ChatGPT-assisted instruction, X_2 denotes Claude-assisted instruction, and -- denotes the conventional lecture method:

Research Design Layout

Group A: O_1 X_1 O_2

Group B: O_1 X_2 O_2

Group C: O_1 --- O_2

The study covered six weeks: Week 1 was for the administration of the pretest across the three groups; Weeks 2 to 5 constituted the treatment period during which students in each group were taught four selected 200 Level computer science concepts, namely Number System Conversion, Flowchart and Algorithm Design, Basic Programming Constructs (variables, conditionals, and loops), and Fundamentals of Computer Networking, using the instructional method assigned to their group, and Week 6 was for the administration of the posttest. The population comprised all Computer Science students in the Colleges of Education in Kwara State, from which three intact classes, one per college, were purposively drawn and assigned to Group A (ChatGPT-assisted), Group B (Claude-assisted), and Group C (conventional method), based on comparable curriculum content, students' baseline familiarity with computer systems, and the availability of functional computer laboratories with adequate internet access across the three colleges, factors considered essential for ensuring that observed differences in academic performance could be reasonably attributed to the instructional method rather than to disparities in digital access or technological exposure among the groups.

College-level differences are confounded with treatment and could not be separated statistically because intact classes from three different colleges were used, with one college per instructional condition. To reduce this threat, the colleges were purposively selected for comparable curriculum content, students' baseline familiarity with computer systems, and functional computer laboratories with internet access. Group equivalence was then checked on the pretest, and pretest scores were entered as a covariate in the

ANCOVA. A one-way ANOVA on the pretest scores showed no significant difference among the three groups, $F(2, 90) = 1.22$, $p = .30$, and the groups did not differ significantly in age, $F(2, 90) = 1.34$, $p = .27$.

Instrument, Validation, Reliability, and Data Analysis

The main instrument for data collection was a 40-item researcher-designed Multiple Choice Performance Test (MCPT), constructed to cover the four selected computer science concepts and administered as a pretest and, with item sequence and answer-option positions altered to reduce rote recall, as a posttest to the three groups. The instrument was subjected to face and content validation by experts in Computer Science Education and Measurement and Evaluation, who assessed its clarity, relevance, and alignment with the course content and instructional objectives, and their corrections were effected before trial testing. Reliability was established through a pilot test administered to 30 NCE II Computer Science students outside the sampled colleges but with similar characteristics. The reliability coefficient of 0.76 was obtained for the MCPT responses using Kuder-Richardson (KR-20). Data collected were analyzed using Analysis of Covariance (ANCOVA), with pretest scores as a covariate, to test the three hypotheses at the 0.05 level of significance. Mean and standard deviation scores were used to answer the research questions and to show the direction of any significant difference found.

RESULTS

Demographic Characteristics of Respondents

A total of 93 NCE II Computer Science students participated in the study, distributed across the ChatGPT-assisted (n = 31), Claude-assisted (n = 33), and Conventional (n = 29) groups. As shown in Table 1, gender distribution was fairly balanced (50.5% male, 49.5% female). Participants' ages ranged from 19 to 27 years ($M = 22.91$, $SD = 2.63$), with comparable average ages across the three groups, suggesting the groups were reasonably equivalent on these background characteristics before treatment.

Table 1: Demographic Distribution of Respondents by Group

Variable	ChatGPT-assisted (n=31)	Claude- assisted (n=33)	Conventional (n=29)	Total (N=93)
Gender: Male	15 (48.4%)	17 (51.5%)	15 (51.7%)	47 (50.5%)
Gender: Female	16 (51.6%)	16 (48.5%)	14 (48.3%)	46 (49.5%)
Age (years): M (SD)	22.29 (2.28)	23.18 (2.88)	23.28 (2.64)	22.91 (2.63)
Age range	19-27	19-27	19-27	19-27

Answering the Research Questions

Research Questions 1 and 2 asked about the effect of ChatGPT-assisted and Claude-assisted instruction on academic performance, while Research Question 3 asked about the difference between the two treatment groups. Table 2 presents the pretest and posttest mean scores. Both AI-assisted groups recorded higher mean gains (ChatGPT: 10.71; Claude: 12.27) than the Conventional group (5.34), with the Claude-assisted group recording the highest posttest mean (26.82) and mean gain, followed by the ChatGPT-assisted group (25.10), and then the Conventional group (18.48). This pattern is illustrated in Figure 1.

Table 2: Pretest and Posttest Scores by Group

Group	N	Pretest M (SD)	Posttest M (SD)	Mean Gain (SD)
ChatGPT-assisted	31	14.39 (4.03)	25.10 (5.21)	10.71 (3.17)
Claude-assisted	33	14.55 (3.90)	26.82 (5.74)	12.27 (3.23)
Conventional	29	13.14 (3.52)	18.48 (5.08)	5.34 (2.53)

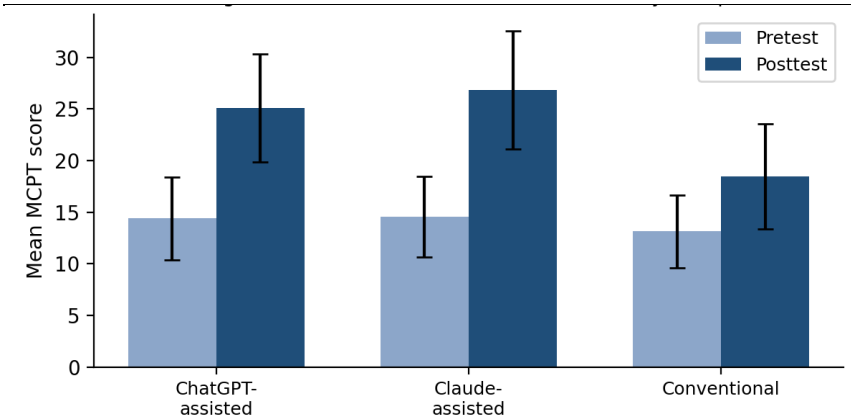


Figure 1: Mean Pretest and Posttest Scores by Group

Test of Hypotheses

A single ANCOVA model, with pretest score as covariate, was fitted using data from all 93 participants to test the overall effect of instructional method on posttest performance (Table 3). Pairwise comparisons of the covariate-adjusted group means were then derived from this same model to test each hypothesis individually (Tables 4–6). This approach, rather than fitting three separate two-group models, was adopted because it draws on the full sample to estimate the pretest–posttest relationship and error variance, yielding more stable and directly comparable estimates across the three pairwise comparisons. Instructional method had a statistically significant effect on posttest performance, $F(2, 89) = 41.44$, $p < .001$, while the covariate (pretest score) was also a significant predictor of posttest performance, $F(1, 89) = 207.99$, $p < .001$, confirming the appropriateness of controlling for it. Each hypothesis was then tested separately using the covariate-adjusted (estimated marginal) means, as follows.

Table 3: Overall ANCOVA Summary for Posttest Scores Controlling for Pretest

Source	SS	Df	MS	F	p
Pretest (covariate)	1814.45	1	1814.45	207.99	<.001
Group (Method)	723.04	2	361.52	41.44	<.001
Error	776.41	89	8.72		

The effect of instructional method was large, $\eta p^2 = .48$, meaning that about 48% of the variance in posttest scores not explained by the pretest was associated with instructional method.

Hypothesis One (H_{01})

H_{01}: There is no significant difference in the academic performance of students taught using ChatGPT-assisted instruction and those taught using the conventional method.

Table 4: ChatGPT-assisted and Conventional

Group	Adjusted Mean	Mean Diff.	F(1,89)	p
ChatGPT-assisted	24.71	5.15	44.76	<.001
Conventional	19.56	-	-	-

The adjusted posttest mean of the ChatGPT-assisted group (24.71) was significantly higher than that of the Conventional group (19.56), F (1, 89) = 44.76, $p < .001$. Since $p < .05$, H_{01} is rejected. ChatGPT-assisted instruction had a significant positive effect on students' academic performance compared with the conventional method.

Hypothesis Two (H_{02})

H_{02}: There is no significant difference in the academic performance of students taught using Claude-assisted instruction and those taught using the conventional method.

Table 5: Claude-assisted and Conventional

Group	Adjusted Mean	Mean Diff.	F(1,89)	P
Claude-assisted	26.24	6.69	77.30	<.001
Conventional	19.56	-	-	-

The adjusted posttest mean of the Claude-assisted group (26.24) was significantly higher than that of the Conventional group (19.56), F(1, 89) = 77.30, $p < .001$. Since $p < .05$, $H0_2$ is rejected. Claude-assisted instruction had a significant positive effect on students' academic performance compared with the conventional method.

Hypothesis Three (H_{03})

H_{03}: There is no significant difference between the mean academic performance scores of students taught using ChatGPT-assisted instruction and those taught using Claude-assisted instruction.

Table 6: Claude-assisted and ChatGPT-assisted

Group	Adjusted Mean	Mean Diff.	F(1,89)	p
Claude-assisted	26.24	1.54	4.32	.041
ChatGPT-assisted	24.71	-	-	-

The adjusted posttest mean of the Claude-assisted group (26.24) was significantly higher than that of the ChatGPT-assisted group (24.71), F(1, 89) = 4.32, $p = .041$. Since $p < .05$, H_{03} is rejected. Claude-assisted instruction produced significantly higher academic performance than ChatGPT-assisted instruction.

Table 7. Effect Sizes for Pairwise Comparisons

Comparison	Adjusted mean diff. [95% CI]	Diff. as % of 40-item test	Cohen's d [95% CI]	ηp²	Interpretation
ChatGPT - Conventional	5.15 [3.63, 6.67]	12.9	1.74 [1.14, 2.35]	.34	Large
Claude - Conventional	6.69 [5.20, 8.18]	16.7	2.27 [1.62, 2.91]	.46	Large
Claude – ChatGPT	1.54 [0.07, 3.01]	3.9	0.52 [0.02, 1.03]	.05	Medium (small to large range)

Table 7 shows the size of each pairwise difference. ChatGPT-assisted instruction exceeded conventional instruction by 5.15 points (95% CI [3.63, 6.67]), about 12.9% of the 40-item test, with a large effect ($d = 1.74$, $\eta p^2 = .34$). Claude-assisted instruction exceeded conventional instruction by 6.69 points (95% CI [5.20, 8.18]), about 16.7% of the test, also with a large effect ($d = 2.27$, $\eta p^2 = .46$). In both cases the confidence intervals were well above zero, so the advantage of AI-assisted instruction over conventional teaching was large and consistently estimated. The difference between Claude-assisted and ChatGPT-assisted instruction was 1.54 points (95% CI [0.07, 3.01]), about 3.9% of the test, with a medium effect ($d = 0.52$, $\eta p^2 = .05$). The interval for this comparison was wide and only just excluded zero, so the advantage of Claude over ChatGPT is small in practical terms and imprecisely estimated, and should be interpreted with caution.

Summary of Findings

1. ChatGPT-assisted instruction produced significantly higher academic performance than the conventional method.
2. Claude-assisted instruction produced significantly higher academic performance than the conventional method.
3. Claude-assisted instruction produced significantly higher academic performance than ChatGPT-assisted instruction, indicating that, within this study, Claude-assisted instruction was associated with the largest performance gains among the three instructional conditions.

DISCUSSION, CONCLUSION, RECOMMENDATIONS AND LIMITATIONS

Discussion of Findings

The finding that ChatGPT-assisted instruction produced significantly higher academic performance than the conventional method aligns with the broader pattern in the literature showing that generative AI can outperform traditional lecture-based teaching (Ukume et al., 2026; Bolarinwa et al., 2026; Elhag et al., 2025). Bolarinwa et al. (2026), for instance, reported significant gains in thermodynamics achievement among students exposed to ChatGPT-assisted instruction compared to a conventional group, a pattern the present study now extends to computer science concepts such as number system conversion and basic programming constructs. This current study, however, opposes the study of Kosar et al. (2024) and Yang et al. (2025), who found no benefit, or even a decline in learning outcomes. A plausible explanation for this divergence is implementation: in the present study, ChatGPT use was structured around four clearly defined computer science concepts over a sustained four-week treatment period, rather than left to unguided student discretion, which is consistent with the view that AI's instructional value depends heavily on the subject area, intervention length, and manner of incorporation into instruction rather than being an automatic or universal effect of the tool itself (Wu et al., 2026; Li et al., 2026).

The finding that Claude-assisted instruction significantly outperformed the conventional method adds new evidence on the value of Claude as a direct teaching tool, an area where research is still limited. Fan et al.

(2025) previously linked Claude 3 Opus to improved programming outcomes, but their students worked in pairs, making it unclear whether the improvement came from Claude itself or simply from having a study partner. The present study tests Claude-assisted instruction as its own distinct treatment, which suggests the gains are associated with Claude itself, although the college confound means this should be interpreted cautiously. Divekar et al. (2024) and Anthropic (2025) had earlier suggested that Claude's guided-reasoning style, asking students how they would approach a problem instead of giving direct answers, may encourage independent thinking rather than replace it. However, that suggestion was based on opinions and exploratory observations, not actual performance data. The present study provides that missing evidence, showing through measured test scores that this instructional style does, in fact, lead to stronger student performance.

The most notable finding is that Claude-assisted instruction produced significantly higher performance than ChatGPT-assisted instruction. At first glance, this seems to disagree with Bernik et al. (2025), who found Claude 3 to be a weaker and less consistent grading tool compared to ChatGPT-4. However, the two findings are not really in conflict. Bernik et al. tested how well Claude scored work that students had already completed, while the present study tested how well Claude taught students in real time. These are different tasks, and Claude's same core trait, prompting students to reason through problems instead of giving direct answers, may explain both results. When used for grading, this trait can produce less standardized, less predictable scores. But when used for teaching, that same trait encourages deeper thinking, which can lead to stronger learning outcomes. This may explain why Claude performed especially well in this study, since concepts like flowchart design, programming constructs, and number systems all depend on step-by-step reasoning (Li et al., 2026).

The effect sizes also help to judge how meaningful these differences are. The advantage of both AI-assisted conditions over conventional teaching was large in practical terms, representing a substantial share of the test score and not a trivial gain. This is consistent with the support and engagement account in the Theoretical Framework, in which on-demand, responsive support from an LLM helps students work through step-by-step content that fixed-pace lectures cannot adapt to individual learners. By contrast, the difference between Claude and ChatGPT was modest and imprecisely estimated, with a wide confidence interval that only just excluded zero. It should therefore be treated as tentative, a possible advantage, and not as evidence that one tool is clearly better than the other.

Conclusion

ChatGPT-assisted and Claude-assisted instruction produced significantly greater academic performance gains among Computer Science students in the colleges of education than the conventional lecture method, and Claude-assisted instruction produced significantly greater gains than ChatGPT-assisted instruction. These results suggest that the choice of LLM is not incidental to instructional outcomes, and that Claude's guided-reasoning design may confer a measurable advantage over ChatGPT's more direct answer-giving style for teaching structured, logic-based computer science content in this population.

Recommendations

i. Computer science lecturers in Colleges of Education should incorporate AI-assisted instruction, particularly Claude, as a supplementary teaching strategy.
ii. Institutional authorities should provide reliable internet access and functional computer laboratories to support AI-assisted instruction, and should organize training for lecturers on how to structure AI-assisted lessons rather than leaving tool use unguided.
iii. Curriculum planners should consider embedding structured AI interactions into computer science instructional design, rather than treating AI tools as interchangeable or leaving their instructional integration informal.

iv. Further studies should replicate this comparison across other computer science concepts, longer treatment periods, and other institutions, and should include a delayed posttest to assess retention beyond immediate performance tests.

Limitations of the Study

i. A quasi-experimental design using intact classes, rather than random assignment of individual students, was adopted for practical and ethical reasons; each treatment group was drawn from a different College of Education, so college-level differences cannot be fully separated from treatment effects despite efforts to select comparable colleges.
ii. The treatment period was limited to four weeks of instruction across only four computer science concepts, which may not capture longer-term or cumulative effects of AI-assisted instruction.
iii. Performance was assessed using an immediate posttest only; without a delayed retention test, it is not possible to determine whether the observed gains, particularly the advantage of Claude-assisted over ChatGPT-assisted instruction, persist over time.
iv. A novelty effect associated with using unfamiliar AI tools cannot be ruled out, as students' engagement may partly reflect the newness of the technology rather than its instructional design alone.

References

Anthropic. (2025, April 2). *Introducing Claude for Education*. https://www.anthropic.com/news/introducing-claude-for-education

Bernik, A., Radošević, D., & Čep, A. (2025). A comparative study of large language models in programming education: Accuracy, efficiency, and feedback in student assignment grading. *Applied Sciences, 15*(18), Article 10055. https://doi.org/10.3390/app151810055

Bolarinwa, O. T., Abubakar, A., Oluwatuberu, A. O., Akintokun, O. A., & Ogungbile, D. O. (2026). Effect of ChatGPT on the academic performance of chemistry students in thermodynamics: A case study of Federal College of Education (Special), Oyo. *International Journal of Research and Scientific Innovation, 13*(3), 1183–1191. https://doi.org/10.51244/ijrsi.2026.1303000111

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Divekar, R. R., Li, H., Xu, Q., & Hebert, C. (2024). *LLMs as academic reading companions: Extending HCI through synthetic personae*. arXiv. https://arxiv.org/abs/2403.19506

Elhag, A., Abri, M. A., & Yousef, A. M. F. (2025). The effect of generative AI tools (ChatGPT, Gemini, etc.) on students' achievement and their motivation towards learning. *Journal of Technology and Science Education*, *15*(3), 746. https://doi.org/10.3926/jotse.3410

Fan, G., Liu, D., Zhang, R., & Pan, L. (2025). The impact of AI-assisted pair programming on student motivation, programming anxiety, collaborative learning, and programming performance: a comparative study with traditional pair programming and individual approaches. *International Journal of STEM Education*, *12*(1). https://doi.org/10.1186/s40594-025-00537-3

Kosar, T., Ostojić, D., Liu, Y. D., & Mernik, M. (2024). Computer science education in ChatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics, 12*(5), Article 629. https://doi.org/10.3390/math12050629

Li, H., Xiao, L., Lu, K., Li, D., Zhang, Z., & Xia, Q. (2026). Exploring the application of large language models (LLMs) in data structure instruction: An empirical analysis of student learning outcomes in computer science. *Information, 17*(4), Article 353. https://doi.org/10.3390/info17040353

Mavrych, V., Yaqinuddin, A., & Bolgova, O. (2025). Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. *Advances in Physiology Education, 49*(2), 430–437. https://doi.org/10.1152/advan.00093.2024

Mienye, I. D., Jere, N., Obaido, G., Ogunruku, O. O., Esenogho, E., & Modisane, C. (2025). Large language models: An overview of foundational architectures, recent trends, and a new taxonomy. *Discover Applied Sciences, 7*(9), Article 1027. https://doi.org/10.1007/s42452-025-07668-w

Mohammed, I. A., Bello, A., & Ayuba, B. (2025). Effect of large language models artificial intelligence ChatGPT chatbot on achievement of computer education students. *Education and Information Technologies, 30*(9), 11863–11888. https://doi.org/10.1007/s10639-024-13293-8

Santos, R. P. D. (2023). *Enhancing chemistry learning with ChatGPT, Bing Chat, Bard, and Claude as agents-to-think-with: A comparative case study*. arXiv. https://arxiv.org/abs/2311.00709

Suleiman, A. A., Aliyu, A., & Mohammed, I. F. (2026). Impact of Gen-AI tools on learning outcomes among students of College of Education Waka-Biu, Borno State Nigeria. *Gombe Journal of Education, 9*(1), 103–114. https://gsujournals.com.ng/gomje/index.php/gomje/article/download/382/210

Ukume, G. D., Jacob, K. J., Uguma, V. U., Okoh, A. S., Egbai, P. O., & Akpa-Inyang, F. F. (2026). Enhancing academic outcomes of year 3 English Education Lexicology students in the University of Calabar, Cross River State, Nigeria using ChatGPT AI: A quasi-experimental study. *International Journal of Learning, Teaching and Educational Research, 25*(1), 779–803. https://www.ijlter.net/index.php/ijlter/article/view/2686

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Vygotsky, L. S. (1978). Mind in Society: Development of Higher Psychological Processes. In *JSTOR*. Harvard University Press. https://www.jstor.org/stable/j.ctvjf9vz4

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications, 13*(1), Article 684. https://doi.org/10.1057/s41599-026-07019-z

Yang, T., Hsu, Y., & Wu, J. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: Evidence from a quasi-experimental research. *Interactive Learning Environments, 33*(6), 3726–3743. https://doi.org/10.1080/10494820.2025.2450659