International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparative Study of Reliability, Bias, and Learning Gains in Claude and chatGPT in Higher Education

Pallavi Chopra[1], Pallavi Sood[2]

[1,2]Assistant Professor, Department Of Computer Science, Apeejay College Of Fine Arts, Jalandhar, Punjab, India

Abstract

Generative artificial intelligence is fundamentally reshaping contemporary instructional structures. This study provides a direct empirical comparison between Claude and ChatGPT to analyze their viability as virtual teaching assistants. By investigating output precision, systemic bias, and student performance metrics, this research reveals distinct behavioral configurations. Claude demonstrates higher factual accuracy and lower demographic bias within STEM subjects. Conversely, ChatGPT proves more effective in humanities instruction due to its narrative fluency and expressive delivery. Notably, while both systems accelerate short-term information retrieval, an over-reliance on automated tools correlates with diminished long-term analytical resilience and critical problem-solving skills. Based on these outcomes, we offer actionable strategic frameworks for administrators and educators to ensure responsible AI integration.

Keywords: Generative AI, Educational Technology, Large Language Models, STEM Education, ChatGPT, Claude, Algorithmic Bias, Student Outcomes, Curriculum Design

1. Introduction

The rapid evolution of large language models (LLMs) has disrupted traditional educational methodologies, introducing novel opportunities for individualized tutoring and interactive academic support . Conversational systems like OpenAI's ChatGPT and Anthropic's Claude are extensively utilized by university cohorts to simplify dense text, conduct literature reviews, and comprehend intricate theoretical concepts. Data from the 2025 EDUCAUSE research report highlights this trend, indicating that over 60% of higher education students across North America and Europe routinely integrate conversational AI into their study routines . Given this widespread adoption, conducting empirical assessments to quantify the educational impacts of these systems is vital. To bridge this gap, this study offers a structural, comparative analysis evaluating the instructional utility of Claude (v4.5 Sonnet) against ChatGPT (v4o).

1.1 Research Objectives

The primary goals of this empirical study are to:

1. **Assess Cross-Disciplinary Precision:** Quantify and contrast the factual truthfulness of ChatGPT and Claude across diverse academic domains.

2. **Examine Algorithmic Skew:** Evaluate and map systemic cultural or demographic partiality embedded within the generated text.
3. **Quantify Learning Benchmarks:** Measure how the integration of automated tutoring assistants affects student performance and mastery of materials.

2. Literature Review

2.1 Algorithmic Pedagogical Assistants

Spurred by transformer advancements, automated learning tools have progressed from rigid, rule-governed platforms like MATHia into highly adaptive, conversational interfaces[1]. These contemporary models foster self-directed learning environments by offering customized dialogues and ongoing context tracking. Existing literature presents conflicting findings regarding their absolute efficacy. For instance, research demonstrates that ChatGPT can match expert benchmarks on specialized professional examinations[2]. However, its operational dependability remains restricted due to the persistent emergence of factual hallucinations.

2.2 Investigating Algorithmic Disparities in Language Models

Scholars have extensively documented embedded bias within natural language processing architectures . Early word-embedding strategies frequently reproduced gendered and racial associations, and systematic linguistic biases against non-standard dialects like African-American English remain prevalent[3]. Within educational environments, these shortcomings pose severe challenges: biased outputs can reinforce cultural clichés and alienate underrepresented student demographics. Although Anthropic attempted to curb these flaws using its "Constitutional AI" paradigm, empirical validations of its real-world success are limited. Furthermore, recent evaluations indicate that GPT-4 still defaults to gender-stereotyped narratives when tasked with generating STEM content[4].

2.3 Educational Impact and System Adoption

AI-mediated learning generally improves immediate conceptual understanding and self-reported academic progress. However, long-term knowledge retention frequently drops below levels achieved via conventional teaching methods[5]. Under the Technology Acceptance Model (TAM), user adoption is primarily driven by perceived system utility and ease of operation. Consequently, analyzing user attitudes remains crucial for evaluating modern educational software.

3. Research Design

This study adopted a concurrent mixed-methods design incorporating both quantitative benchmarks and qualitative feedback.

3.1 Phase 1 – Evaluation of Correctness

An evaluation dataset consisting of 600 standardized academic questions across Mathematics, Physics, Computer Science, History, Literature, and Economics was compiled from verified repositories, including the MMLU[6]. Both language models were tested using zero-shot prompting strategies. Subsequently, three independent expert reviewers graded the outputs on a 5-point Likert scale(1=Poor, 5= Excellent) to assess factual accuracy, content thoroughness, and instructional clarity. The inter-rater consistency was verified as robust, yielding a Cohen's Kappa of 0.76. The outcome generated in comparative performance analysis of two models across evaluation criteria.

3.2 Phase 2 – Assessment of Prejudice

To assess systemic bias, researchers designed 200 tailored situational prompts containing distinct cultur-

al markers, naming conventions, and contextual backdrops. Prompts are submitted to the target language model(s) to generate responses[7] . analytical pipeline (Multi-method approach) combined automated sentiment tracking (using VADER and BERT architectures), LDA topic modeling to identify dominant themes and discourse patterns, and manual stereotype classification to classify responses for presence of stereotypes. Algorithmic bias was quantified by calculating the mean absolute deviation across the investigated demographic variations.

n

Bias(Mean Absolute Deviation) across demographic groups= MAD= $1/n \sum |x_i - \bar{x}|$

i=1

- x_i = bias score for demographic group i
- $\bar{x}$ = mean bias score across all groups
- n= number of demographic groups

3.3 Phase 3 – Monitored Educational Trial

A cohort of 120 undergraduate students was randomly allocated to one of three distinct educational environments:

- **Group A:** Learning assisted exclusively by Claude .
- **Group B:** Learning assisted exclusively by ChatGPT.
- **Group C:** Traditional independent study without AI tools.

The instructional materials covered Data Structures and Modern World History. Pre-test (Baseline assessment of knowledge) and post-test (Assessment of learning outcomes) instruments measured net learning gains, with an ANCOVA model utilized to control for baseline GPA variations[8]. Qualitative engagement patterns were derived from student reflection journals.

4. Findings

4.1 Accuracy and Reliability Metrics

Statistical testing showed that Claude maintained a slight advantage in overall accuracy, scoring an average of 4.21 out of 5 compared to ChatGPT's 4.07 out of 5, with Claude exhibiting a prominent lead in STEM fields. Conversely, ChatGPT excelled in humanities courses, benefiting from a highly articulate delivery and expansive contextual framing . Regarding factual errors, Claude maintained a 6.2% hallucination rate, whereas ChatGPT reached 8.7%. This disparity became more pronounced during technical queries, where ChatGPT's inaccuracy rate rose to 7.9%, nearly double Claude's 4.1%.

Figure 1 : Accuracy and Reliability Metrics: Claude Vs ChatGPT

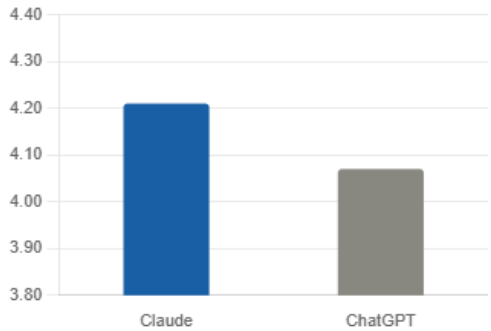

Fig. 1.1: Overall accuracy score (Out of 5)

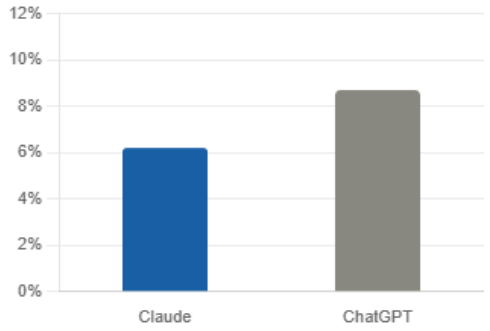

Fig. 1.2: Hallucination Rate (%)

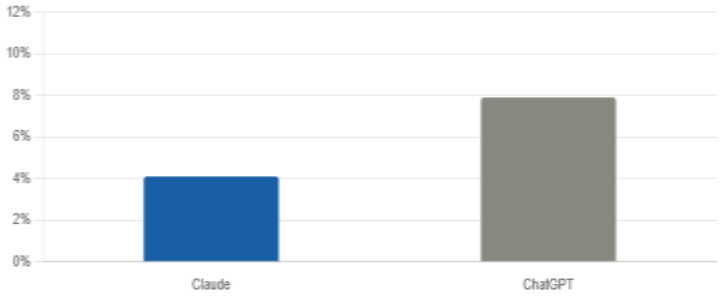

Fig. 1.3: Inaccuarcy rate on Technical queries (%)

Note: Higher accuracy scores are better, lower hallucination and inaccuracy rates are better

4.2 Systemic Bias Metrics

In bias evaluations, Claude demonstrated a lower index score (0.155) than ChatGPT (0.248), a margin established as statistically significant at ($p < 0.01$\). A shared Eurocentric orientation was apparent in both systems. Biased or stereotypical outputs were identified in 7.2% of Claude's generation dataset, compared to 11.4% within ChatGPT's processed content.

Figure 2: Bias Evaluation Metrics: Claude Vs ChatGPT

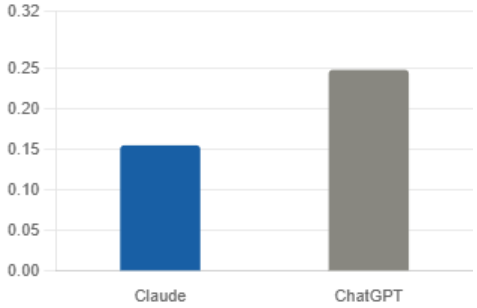

Fig. 2.1 Bias index score (lower= less biasesd)

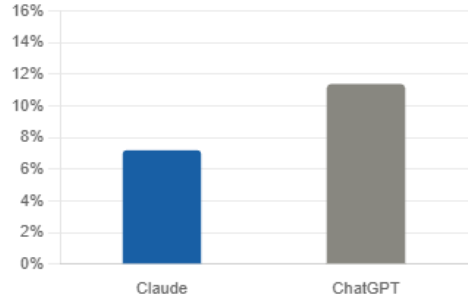

Fig. 2.2 Biased/ Stereotypical output (%)

Note: Bias index difference is statistically significant ($p<0.01$). Both system exhibited a shared Eurocentric orientation.

4.3 Student Knowledge Growth

Baseline diagnostics confirmed that all student cohorts began with equivalent domain knowledge ($p = 0.87$). Post-intervention assessments revealed that both AI-assisted pathways yielded significant intellectual growth over the traditional study group ($p < 0.001$):

- **Claude Cohort:** +17.5% score increase
- **ChatGPT Cohort:** +15.5% score increase
- **Non-AI Control Cohort:** +8.1% score increase

While the post-test score differences between Claude and ChatGPT users were not statistically meaningful ($p = 0.08$) , student journals exposed diverging educational dynamics. ChatGPT acted as a catalyst for inquiry and interactive debate, while Claude proved more effective at establishing student self-assurance in complex academic ideas.

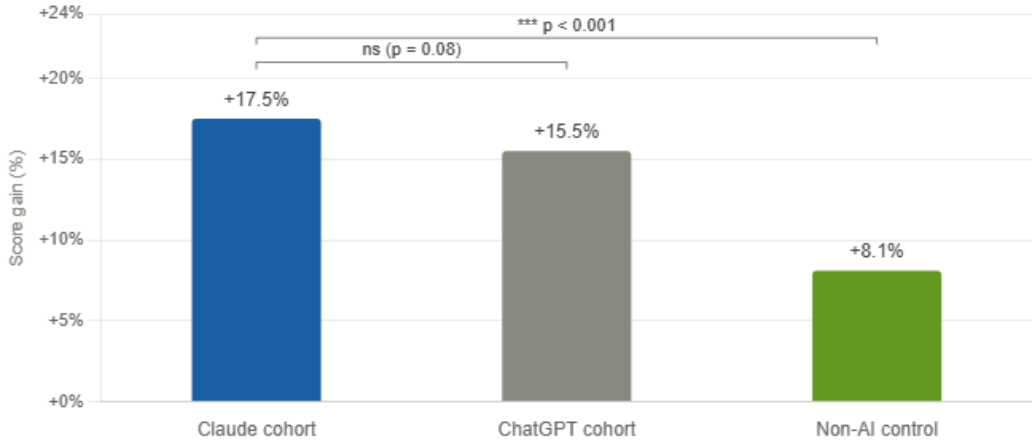

Fig.3 Post-intervention score gains by Cohort (Baseline equivalence confirmed: p=0.87(ns))

Note: p<0.001(AI-assisted Vs non-AI control, ns= not significant (p=0.08, Claude Vs ChatGPT ChatGPT promoted inquiry and debate while Claude strengthened student's confidence in complex concepts.

5. Discussion

5.1 Precision Variances and Pedagogical Delivery

The superior factual reliability shown by Claude corresponds with its structured, conservative reasoning framework[9]. Conversely, ChatGPT's vibrant and articulate output offers distinct advantages in conversational or debate-driven courses. Academic designers should balance these unique attributes—deploying Claude for rigid, fact-based curriculum and ChatGPT for fluid, interpretive analysis—while enforcing instructor oversight to halt the spread of automated errors.

5.2 Cultural Perspectives and Fair Representation

Both language models replicated subtle Eurocentric viewpoints, mirroring the systemic cultural imbalances embedded within modern web-scale training data. Even though Claude demonstrated fewer instances of prejudice, the remaining fluctuations emphasize the necessity for culturally diverse prompt framing and localized balance in educational examples[10]. Introducing explicit critique of these algorithmic biases in classrooms can protect against epistemic imbalances.

5.3 Cognitive Outcomes of Automated Tutoring

The pronounced immediate academic improvements confirm the practical value of AI integration. However, student diaries revealed that nearly 20% of participants reduced their independent reading efforts, a behavior that reflects established cognitive offloading tendencies[11]. Excessive reliance on AI assistants may compromise the development of independent analytical skills, which highlights the critical need for multi-year tracking studies [12].

5.4 User Engagement and Trust Frameworks

Diverging student feedback highlights a clear trade-off between intuitive usability and analytical clarity. While ChatGPT's fluid conversational interface elevates learner motivation, Claude's careful, methodical reasoning instills deeper user trust [13]. Implementing integrated learning frameworks that merge both platforms could simultaneously optimize student motivation and cognitive development[14].

6. Conclusions and Recommendations

This comparative investigation provides the initial cohesive evaluation of Claude and ChatGPT as virtual teaching assistants. The primary findings indicate the following:

- **System Dependability:** Claude exhibits superior factual precision and minimal socio-demographic skew, which is particularly evident in science and technology fields.
- **Narrative Expressiveness:** ChatGPT provides more engaging storytelling and emotional connection, making it highly effective for arts and humanities education.
- **Educational Impact:** Both large language models successfully accelerate short-term student progress when compared to traditional, non-AI learning techniques.
- **Caution:** The presence of geographic data bias alongside the risk of diminishing students' independent critical thinking means that instructors must provide deliberate, proactive supervision.

6.1 Policy Recommendations

1. **Subject-Tailored Tool Deployment:** Align specific generative models with the unique demands of academic fields. Deploy Claude for structured, quantitative tasks and ChatGPT for open-ended, fluid discussions.
2. **Instructional Oversight:** Integrate proactive educator intervention, such as fact-checking assignments, group debates, and self-reflection exercises, to preserve student critical thinking.
3. **Supplemental Integration:** Position artificial intelligence platforms strictly as teaching assistants designed to support learning, rather than full substitutes for foundational curriculum materials.
4. **Diverse Representation:** Proactively introduce varied cultural markers, perspectives, and cross-cultural examples into system prompts to expand inclusivity.
5. **Long-Term Research Initiatives:** Launch multi-year, multi-language research projects to monitor the lasting cognitive developments and ethical impacts of automated learning.

References

1. Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI feedback*. Anthropic Research Preprint.
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM FAccT Conference*, 610–623.

3. Blodgett, S. L., Barocas, S., Daumé, H. III, & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the ACL*, 5454–5474.
4. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 4349–4357.
5. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
6. EDUCAUSE. (2025). *Higher education and AI adoption survey report*. EDUCAUSE Research Publications .
7. Hendrycks, D., Burns, C., et al. (2021). Measuring massive multitask language understanding (MMLU). *arXiv preprint arXiv:2009.03300*.
8. Kung, T. H., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198.
9. Mollick, E. R., & Mollick, L. (2023). Assigning ChatGPT: Student perception and outcomes. *Wharton School Working Paper Series*.
10. Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as socio-technical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684.
11. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
12. Shaikh, S., et al. (2023). Bias persistence in generative AI's educational use: A critical assessment. *Educational Review*, 75(4), 512–531.
13. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
14. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.