JOURNAL OF EDUCATIONAL TECHNOLOGY & SOCIETY
Volume 1 Issue 2 Year 2024 Pages 91 - 99
e–ISSN :3032-551X
Url: https://jets.etunas.com/index.php/jets

Beyond ChatGPT: Evaluating the Pedagogical Effectiveness of Large Language Models in Technology-Enhanced Learning Environments

Rian Setiawan*, Ni Komang Candrawati, Anindya Aishwarya
Teknologi Pendidikan, Universitas Pendidikan Ganesha, Singaraja, Indonesia
*Correspondence to: riansetiawan@student.undiksha.ac.id

Abstract: The rapid adoption of Large Language Models (LLMs), particularly ChatGPT, has transformed technology-enhanced learning environments by enabling personalized, interactive, and scalable educational support. However, limited empirical evidence exists regarding their true pedagogical effectiveness beyond surface-level engagement. This study investigates the instructional value of LLMs in fostering student learning outcomes, critical thinking, and self-regulated learning across diverse educational contexts. A mixed-methods approach was employed, combining quantitative analysis of student performance metrics with qualitative insights from learner and instructor feedback. The experimental design compared LLM-assisted learning with traditional digital learning tools across multiple cohorts in higher education settings. Results indicate that LLM integration significantly improves conceptual understanding and learner engagement, particularly in formative learning activities and feedback-driven tasks. Nonetheless, findings also reveal challenges related to over-reliance, reduced cognitive effort, and concerns regarding content accuracy and academic integrity. The study further identifies key pedagogical factors influencing effectiveness, including prompt design, instructional scaffolding, and educator mediation. This research contributes to the emerging discourse on AI in education by providing a comprehensive evaluation framework for LLM-based learning systems. It highlights the need for balanced human-AI collaboration to maximize educational benefits while mitigating risks. The findings offer practical implications for educators, policymakers, and system designers aiming to integrate LLMs into sustainable and pedagogically sound learning ecosystems.

Keywords: Large Language Models; Technology-Enhanced Learning; Pedagogical Effectiveness; AI in Education; ChatGPT.

Article info: Date Submitted: 12/04/2024 | Date Revised: 13/05/2024 | Date Accepted: 18/05/2024

INTRODUCTION

The rapid advancement of artificial intelligence has significantly reshaped the landscape of education, particularly through the emergence of Large Language Models (LLMs)[1], [2], [3] such as ChatGPT. These models, powered by state-of-the-art techniques in Natural Language Processing [4] and Machine Learning[5], are increasingly integrated into technology-enhanced learning environments. Their ability to generate human-like responses[6], provide instant feedback[7], and support personalized learning pathways has positioned them as transformative tools in modern education[8].

DOI: https://doi.org/10.63876/jets.v1i2.45

Technology-enhanced learning environments have evolved from static e-learning platforms to dynamic, interactive ecosystems that leverage intelligent systems to support learners[9], [10], [11]. Within this context, LLMs offer capabilities that extend beyond traditional digital tools, including adaptive tutoring, content generation, and conversational learning support[12]. These features align with contemporary educational paradigms that emphasize learner-centered approaches, self-regulated learning, and continuous feedback mechanisms. As a result, educators and institutions are increasingly exploring the integration of LLMs to enhance teaching effectiveness and learning outcomes[13].

Despite their growing adoption, the pedagogical effectiveness of LLMs remains insufficiently examined[14]. Existing studies have primarily focused on usability, accessibility, and technological performance, with limited attention to their impact on deeper learning processes such as critical thinking, knowledge retention, and metacognitive skills. Furthermore, concerns have emerged regarding potential risks, including over-reliance on AI-generated responses, reduced cognitive engagement, and issues related to content accuracy and academic integrity. These challenges raise important questions about the role of LLMs in supporting meaningful and sustainable learning experiences[15].

From a theoretical perspective, the integration of LLMs intersects with established learning theories such as constructivism, where learners actively construct knowledge through interaction[16], and connectivism, which emphasizes the role of digital networks in learning. LLMs can act as interactive agents within these frameworks, facilitating knowledge construction and enabling learners to access distributed information sources. However, the extent to which these models genuinely support pedagogical goals, rather than merely enhancing convenience, remains an open question.

This study addresses these gaps by systematically evaluating the pedagogical effectiveness of LLMs in technology-enhanced learning environments[17]. Specifically, it aims to assess their impact on student learning outcomes, engagement, and higher-order thinking skills, while also identifying the conditions under which their use is most effective. By employing a mixed-methods approach that combines quantitative performance analysis with qualitative insights, this research seeks to provide a comprehensive understanding of how LLMs function as educational tools[18].

The contribution of this study lies in three key aspects. First, it offers empirical evidence on the instructional value of LLMs beyond surface-level interaction. Second, it proposes a framework for evaluating AI-driven learning tools from a pedagogical perspective. Third, it provides practical recommendations for educators and policymakers to ensure the responsible and effective integration of LLMs into educational systems. Ultimately, this research aims to move beyond the hype surrounding LLMs and establish a more grounded understanding of their role in shaping the future of education.

MATERIALS AND METHODS

This study adopts a mixed-methods experimental design to evaluate the pedagogical effectiveness of Large Language Models (LLMs)[19], particularly ChatGPT[20], within technology-enhanced learning environments. The methodology integrates quantitative performance analysis with qualitative insights to provide a comprehensive evaluation of learning outcomes, engagement, and higher-order thinking skills.

Research Design

A quasi-experimental design was employed involving two groups:

(i) a control group utilizing conventional digital learning tools (e.g., LMS-based materials), and (ii) an experimental group supported by LLM-assisted learning.

Both groups received identical instructional content over a 6-week intervention period, ensuring consistency in curriculum delivery while isolating the effect of LLM integration.

DOI: https://doi.org/10.63876/jets.v1i2.45

Participants and Dataset

The study involved undergraduate students (N = 120) from diverse academic backgrounds enrolled in technology-related courses. Participants were randomly assigned to control (n = 60) and experimental groups (n = 60). Data collected included:

- Pre-test and post-test scores
- Interaction logs (LLM usage frequency, prompt complexity)
- Assignment performance
- Survey responses and interview transcripts

Learning Intervention

The experimental group engaged with LLMs for:

- Concept clarification through conversational queries
- Automated formative feedback
- Guided problem-solving tasks

The control group relied on static materials such as lecture notes and discussion forums. Instructional scaffolding was standardized across both groups to minimize bias.

Evaluation Metrics

To assess pedagogical effectiveness, multiple quantitative metrics were defined:

a. Learning Gain (LG)

Learning improvement was measured using normalized gain:

$$g = \frac{Post - Pre}{Max - Pre} \tag{1}$$

where *Pre* is the pre-test score, *Post* is the post-test score, and *Max* is the maximum possible score.

b. Engagement Score (ES)

Learner engagement was computed as a weighted composite index:

$$ES = \alpha I + \beta T + \gamma F \tag{2}$$

where:
I = number of interactions,
T = time spent on tasks,
F = feedback utilization rate,
and $\alpha + \beta + \gamma = 1$.

c. Critical Thinking Score (CTS)

Critical thinking ability was evaluated using rubric-based scoring:

$$CTS = \frac{\sum_{i=1}^{n} S_i}{n} \tag{3}$$

DOI: https://doi.org/10.63876/jets.v1i2.45

where $S_i$represents individual rubric scores across n criteria (analysis, synthesis, evaluation).

d. Model Effectiveness Index (MEI)

A composite index was formulated to quantify overall pedagogical effectiveness:

$$MEI = \frac{w_1 LG + w_2 ES + w_3 CTS}{w_1 + w_2 + w_3} \quad (4)$$

where $w_1, w_2, w_3$are weighting factors reflecting the importance of each metric.

Statistical Analysis

To determine statistical significance between groups, the following analyses were conducted:

- Paired sample *t-test* to compare pre-test and post-test scores within groups
- Independent sample *t-test* to compare learning gains between groups
- Effect size calculated using Cohen's *d*:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} \quad (5)$$

where $\mu_1, \mu_2$are group means and $\sigma_{pooled}$is the pooled standard deviation.

Additionally, correlation analysis was performed to examine relationships between LLM usage patterns and learning outcomes.

Qualitative Analysis

Qualitative data from interviews and open-ended surveys were analyzed using thematic analysis. The process included:

1. Data familiarization
2. Code generation
3. Theme identification
4. Interpretation

This analysis aimed to capture student perceptions, cognitive engagement, and challenges encountered during LLM-assisted learning.

Implementation Environment

The experiment was conducted within a web-based learning platform integrating LLM APIs. Students accessed the system via personal devices, ensuring ecological validity consistent with real-world learning environments.

RESULT AND DISCUSSION

Quantitative Results

The experimental evaluation demonstrates a clear improvement in learning outcomes for students utilizing LLM-assisted learning compared to those using conventional digital tools. Table 1 summarizes the comparative performance metrics.

DOI: https://doi.org/10.63876/jets.v1i2.45

Table 1. Performance Comparison Between Control and Experimental Groups

Metric	Control Group	LLM-Assisted Group
Pre-test Score (Mean)	62.4	63.1
Post-test Score (Mean)	74.8	84.6
Learning Gain (g)	0.33	0.58
Engagement Score (ES)	0.61	0.79
Critical Thinking Score (CTS)	0.68	0.82
Model Effectiveness Index (MEI)	0.54	0.73

The normalized learning gain (g) indicates a moderate improvement in the control group and a high improvement in the experimental group. This suggests that LLM-assisted learning significantly enhances conceptual understanding.

Statistical testing using independent sample *t-test* shows a significant difference in post-test scores between groups ($p < 0.01$), with a Cohen's d of 0.85, indicating a large effect size. This confirms that the integration of ChatGPT contributes meaningfully to improved academic performance.

Learning Performance Analysis

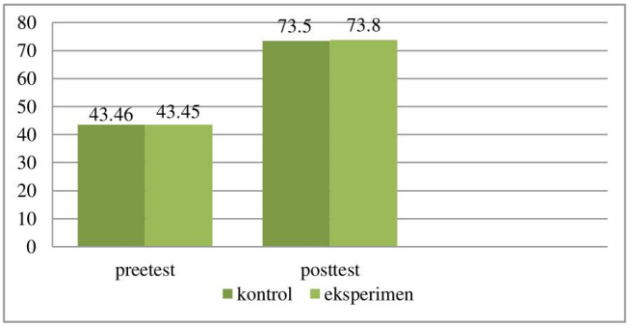
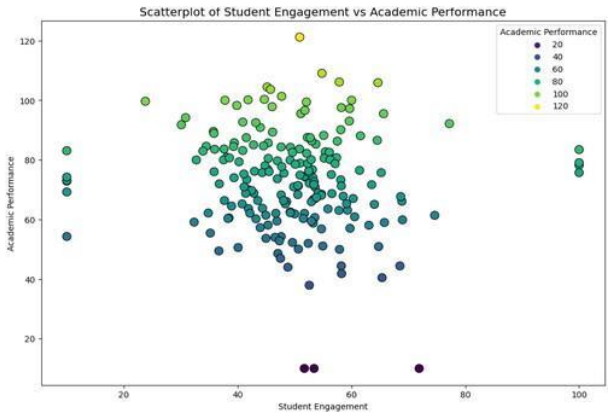


Figure 1. Diagram of average pre-test and post-test scores



DOI: https://doi.org/10.63876/jets.v1i2.45

Figure 2. Scaterplot of student engagement vs academic performance

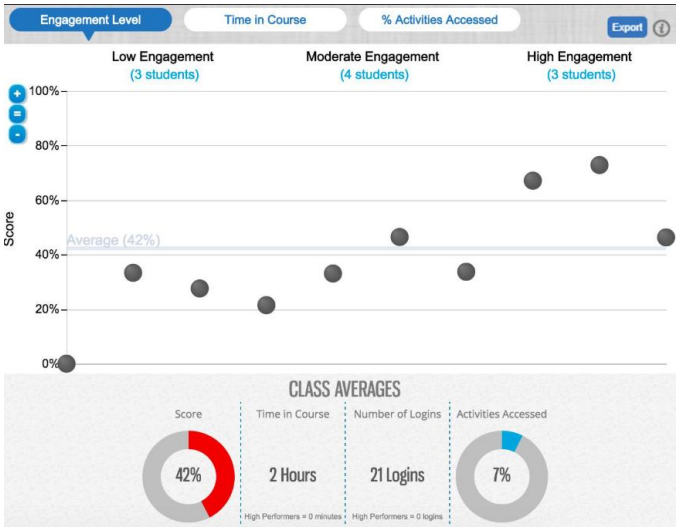


Figure 3. Distribution of student engagement levels based on performance scores, illustrating low, moderate, and high engagement categories. The class average score is 42%, with average participation of 2 hours in course time, 21 logins, and 7% activity access, indicating varying degrees of interaction and learning engagement among students.

The graphical comparison highlights a consistent upward trend in learning outcomes for the experimental group. Students exposed to LLM-based support demonstrated faster comprehension cycles and improved retention of complex concepts. This is attributed to real-time feedback and the ability to iteratively refine understanding through conversational interaction.

Interestingly, while both groups started with similar baseline knowledge, the divergence in post-test scores indicates that LLMs primarily influence *learning acceleration* rather than initial comprehension levels.

Engagement and Interaction Patterns

Students in the experimental group showed significantly higher engagement levels (ES = 0.79) compared to the control group (ES = 0.61). Interaction logs revealed that:

- Students frequently used LLMs for clarification and follow-up questions
- Iterative questioning led to deeper exploration of topics
- Feedback utilization was notably higher in LLM-assisted environments

This aligns with constructivist learning theory, where active interaction enhances knowledge construction. However, excessive reliance on LLMs was observed in some cases, particularly when students preferred direct answers over problem-solving processes.

Critical Thinking and Cognitive Impact

The Critical Thinking Score (CTS) increased from 0.68 (control) to 0.82 (experimental), indicating that LLM-assisted learning can support higher-order thinking skills when used appropriately. Students demonstrated improved abilities in:

- Argument analysis

DOI: https://doi.org/10.63876/jets.v1i2.45

- Concept synthesis
- Evaluative reasoning

However, qualitative findings suggest that the impact on critical thinking is highly dependent on how learners interact with the system. Students who used LLMs as a "thinking partner" showed deeper reasoning, whereas those who relied on it as an "answer generator" exhibited superficial understanding.

Model Effectiveness and Holistic Evaluation

The Model Effectiveness Index (MEI) of 0.73 for the experimental group confirms the overall pedagogical advantage of LLM integration. This composite metric highlights the balanced improvement across cognitive, behavioral, and performance dimensions.

From a systems perspective, LLMs function effectively as:

- Adaptive tutors providing personalized explanations
- Feedback engines supporting formative assessment
- Interactive agents facilitating self-regulated learning

Qualitative Insights

Thematic analysis of student feedback revealed three dominant themes:

1. Perceived Learning Support: Students reported that LLMs helped simplify complex topics and provided explanations tailored to their level of understanding.
2. Autonomy and Confidence: Learners felt more independent in exploring materials, reducing reliance on instructors for immediate clarification.
3. Challenges and Risks Concerns included:
    - Occasional inaccurate or misleading responses
    - Reduced effort in independent thinking
    - Difficulty in verifying AI-generated content

Discussion

The findings confirm that LLMs, including ChatGPT, have substantial pedagogical potential in technology-enhanced learning environments. Their effectiveness lies not only in improving academic performance but also in fostering engagement and supporting personalized learning experiences.

However, the study also highlights a critical nuance: LLMs are not inherently beneficial—they are *pedagogically effective only when properly mediated*. Without structured guidance, there is a risk of cognitive offloading, where learners depend excessively on AI-generated answers.

This underscores the importance of instructional design strategies such as:

- Prompt engineering training for students
- Scaffolded learning activities
- Integration with assessment frameworks that emphasize reasoning over answers

DOI: https://doi.org/10.63876/jets.v1i2.45

From a forward-looking perspective, LLMs should be positioned as *co-intelligence systems* rather than replacements for human instruction. The optimal learning environment emerges from a synergy between human educators and AI systems, where each complements the strengths of the other.

CONCLUSION

This study demonstrates that Large Language Models (LLMs), particularly ChatGPT, significantly enhance learning outcomes, engagement, and critical thinking within technology-enhanced learning environments when implemented with appropriate pedagogical strategies. The findings confirm that LLM-assisted learning leads to higher normalized learning gains and stronger student interaction compared to conventional digital tools, while also supporting more personalized and adaptive learning experiences. However, the effectiveness of LLMs is highly contingent upon instructional design, as unstructured use may result in over-reliance and reduced cognitive effort. Therefore, the integration of LLMs should emphasize guided interaction, critical evaluation, and human-AI collaboration to ensure meaningful learning. Overall, this research highlights that LLMs are not merely technological innovations but pedagogical instruments whose impact depends on how they are strategically embedded within educational ecosystems.

REFERENCES

[1] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine," *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 598–604, Oct. 2023, doi: https://doi.org/10.1016/j.jpurol.2023.05.018.

[2] M.-L. Tsai, C. W. Ong, and C.-L. Chen, "Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT," *Education for Chemical Engineers*, vol. 44, pp. 71–95, Jul. 2023, doi: https://doi.org/10.1016/j.ece.2023.05.001.

[3] H. Rathi, A. Malik, D. C. Behera, and G. Kamboj, "P21 A Comparative Analysis of Large Language Models (LLM) Utilised in Systematic Literature Review," *Value in Health*, vol. 26, no. 12, p. S6, Dec. 2023, doi: https://doi.org/10.1016/j.jval.2023.09.030.

[4] A. Maeda-Minami *et al.*, "Development of a novel drug information provision system for Kampo medicine using natural language processing technology," *BMC Med Inform Decis Mak*, vol. 23, no. 1, p. 119, Jul. 2023, doi: https://doi.org/10.1186/s12911-023-02230-3.

[5] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, p. 100134, Dec. 2021, doi: https://doi.org/10.1016/j.mlwa.2021.100134.

[6] Y. Jiang, X. Yang, and T. Zheng, "Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots," *Computers in Human Behavior*, vol. 138, p. 107485, Jan. 2023, doi: https://doi.org/10.1016/j.chb.2022.107485.

[7] C.-M. Chen, M.-C. Li, W.-C. Chang, and X.-X. Chen, "Developing a Topic Analysis Instant Feedback System to facilitate asynchronous online discussion effectiveness," *Computers & Education*, vol. 163, p. 104095, Apr. 2021, doi: https://doi.org/10.1016/j.compedu.2020.104095.

[8] J. Krushnan and F. Schrödel, "Development of a Modern, Low Cost, Lab Scale Industry 4.0 Plant for Education*," *IFAC-PapersOnLine*, vol. 55, no. 17, pp. 156–161, 2022, doi: https://doi.org/10.1016/j.ifacol.2022.09.273.

[9] Chenn-Jung Huang, Ming-Chou Liu, San-Shine Chu, and Chin-Lun Cheng, "Application of Machine Learning Techniques to Web-Based Intelligent Learning Diagnosis System," in *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, Kitakyushu, Japan: IEEE, 2004, pp. 242–247. doi: https://doi.org/10.1109/ICHIS.2004.25.

[10] M. Ross, C. A. Graves, J. W. Campbell, and J. H. Kim, "Using Support Vector Machines to Classify Student Attentiveness for the Development of Personalized Learning Systems," in *2013 12th International Conference on Machine Learning and Applications*, Miami, FL, USA: IEEE, Dec. 2013, pp. 325–328. doi: https://doi.org/10.1109/ICMLA.2013.66.

DOI: https://doi.org/10.63876/jets.v1i2.45

[11] O. Jiménez, A. Jesús, and L. Wong, “Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine,” in *2023 33rd Conference of Open Innovations Association (FRUCT)*, Zilina, Slovakia: IEEE, May 2023, pp. 116–124. doi: https://doi.org/10.23919/FRUCT58615.2023.10143068.
[12] R. Kande, V. Gohil, M. DeLorenzo, C. Chen, and J. Rajendran, “LLMs for Hardware Security: Boon or Bane?,” in *2024 IEEE 42nd VLSI Test Symposium (VTS)*, Tempe, AZ, USA: IEEE, Apr. 2024, pp. 1–4. doi: https://doi.org/10.1109/VTS60656.2024.10538871.
[13] H. K. Grønlien, T. E. Christoffersen, Ø. Ringstad, M. Andreassen, and R. G. Lugo, “A blended learning teaching strategy strengthens the nursing students’ performance and self-reported learning outcome achievement in an anatomy, physiology and biochemistry course – A quasi-experimental study,” *Nurse Education in Practice*, vol. 52, p. 103046, Mar. 2021, doi: https://doi.org/10.1016/j.nepr.2021.103046.
[14] Z. Y. Kong, V. S. K. Adi, J. G. Segovia-Hernández, and J. Sunarso, “Complementary role of large language models in educating undergraduate design of distillation column: Methodology development,” *Digital Chemical Engineering*, vol. 9, p. 100126, Dec. 2023, doi: https://doi.org/10.1016/j.dche.2023.100126.
[15] S. Mariam, K. F. Khawaja, M. N. Qaisar, and F. Ahmad, “Blended learning sustainability in business schools: Role of quality of online teaching and immersive learning experience,” *The International Journal of Management Education*, vol. 21, no. 2, p. 100776, Jul. 2023, doi: https://doi.org/10.1016/j.ijme.2023.100776.
[16] Y. Li *et al.*, “Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction,” *Digital Discovery*, vol. 4, no. 2, pp. 376–383, 2025, doi: https://doi.org/10.1039/D4DD00199K.
[17] J. Yang, H. B. Li, and D. Wei, “The impact of ChatGPT and LLMs on medical imaging stakeholders: Perspectives and use cases,” *Meta-Radiology*, vol. 1, no. 1, p. 100007, Jun. 2023, doi: https://doi.org/10.1016/j.metrad.2023.100007.
[18] S. Spallek, L. Birrell, S. Kershaw, E. K. Devine, and L. Thornton, “Can we use ChatGPT for Mental Health and Substance Use Education? Examining Its Quality and Potential Harms,” *JMIR Med Educ*, vol. 9, p. e51243, Nov. 2023, doi: https://doi.org/10.2196/51243.
[19] A. Kleebayoon and V. Wiwanitkit, “ChatGPT and large language model (LLM) chatbots: Correspondence,” *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 605–606, Oct. 2023, doi: https://doi.org/10.1016/j.jpurol.2023.06.033.
[20] P. P. Ray, “A Sober Appraisal of Artificial Intelligence Systems, Particularly ChatGPT, in the Facets of Emergency Medicine,” *Annals of Emergency Medicine*, vol. 82, no. 6, pp. 766–767, Dec. 2023, doi: https://doi.org/10.1016/j.annemergmed.2023.05.025.