Pantonic Review
www.patonicreview.org
Peer review, Open Access Journal

CareGenerative AI in Teacher Education: Capabilities and Biases of Large Language Models in Situational Judgment Tests and Pedagogical Decision-Making

Dr. Gufran Alam
Assistant Professor, Women's Training College, Patna University

Abstract

The integration of generative artificial intelligence, particularly Large Language Models (LLMs), into teacher education represents a transformative yet contested development. This article examines the capabilities and biases of LLMs when applied to two critical domains in teacher preparation: situational judgment tests (SJTs) for assessing professional knowledge, and pedagogical decision-making in classroom contexts. Drawing on recent empirical research, we find that LLMs demonstrate significant promise in generating SJT items with psychometric properties comparable to expert-developed assessments, while substantially reducing development time and resource costs. However, LLMs also exhibit systematic biases—including racial, gender, and socioeconomic biases—when evaluating teaching scenarios and generating educational content. These biases manifest across multiple modalities, from text-based evaluations to vision-language models processing classroom imagery. We synthesize findings on LLM capabilities in automated scenario generation, response analysis, and simulation-based training, while critically examining the ethical implications of algorithmic bias in teacher evaluation. The article concludes with recommendations for responsible integration, emphasizing the necessity of human oversight, critical digital pedagogy, and continuous bias auditing in teacher education programs.

Keywords: generative AI, large language models, teacher education, situational judgment tests, pedagogical decision-making, algorithmic bias, fairness

Pantonic Review
Volume 1 | Issue 1 | 2026 | Article ID 150013 | http://doi.org/10.5281/zenodo.23661675
www.pantonicreview.org

© 2026 The Author(s). Published by Pantonic Review. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the license is provided, and any changes are indicated.

1. Introduction

The rapid proliferation of generative artificial intelligence (GenAI) across educational contexts has prompted urgent questions about its role in teacher preparation and professional development. Large Language Models (LLMs)—advanced AI systems trained on vast text corpora—are increasingly being deployed in teacher education for tasks ranging from lesson planning assistance to the automated evaluation of teaching competencies. While these technologies offer unprecedented opportunities for scalable, personalized training and assessment, they also introduce profound challenges related to algorithmic bias, fairness, and the erosion of professional judgment.

Two domains where LLMs are gaining particular traction are situational judgment tests (SJTs) and pedagogical decision-making. SJTs are assessment tools that present educators with realistic classroom scenarios to evaluate their professional knowledge and judgment. Traditionally, SJT development has been labor-intensive, requiring extensive input from subject matter experts. LLMs offer the potential to automate and accelerate this process, generating scenarios and response options with efficiency that far exceeds conventional methods. Concurrently, LLMs are being explored as tools to support pedagogical decision-making simulating classroom interactions, providing feedback to pre-service teachers, and even evaluating teaching quality.

Yet these capabilities are shadowed by significant concerns. Research has documented that LLMs exhibit systematic biases when operating in "teacher" roles, perpetuating stereotypes related to race, gender, disability status, and socioeconomic background. These biases are not mere artifacts but reflect the prejudices embedded in the massive datasets on which LLMs are trained. When deployed in high-stakes educational contexts—such as teacher evaluation or the assessment of professional competence such biases risk amplifying existing inequities rather than mitigating them.

This article aims to provide a comprehensive examination of both the capabilities and limitations of LLMs in teacher education, with specific focus on SJTs and pedagogical decision-making. We synthesize findings from recent empirical studies, identify patterns of bias across different LLM architectures and applications, and propose frameworks for responsible integration. Our analysis is guided by three central questions:

(1) What are the demonstrated capabilities of LLMs in generating and evaluating SJTs for teacher education? (2) What forms of bias manifest when LLMs are applied to pedagogical judgment tasks? (3) How can teacher education programs harness LLM capabilities while mitigating their biases?

http://doi.org/10.5281/zenodo.23661675

2. Large Language Models in Situational Judgment Test Development

2.1 The Promise of Automated SJT Generation

Situational Judgment Tests have long been recognized as valuable tools for assessing professional competencies across fields including medicine, military leadership, and education. SJTs present respondents with realistic scenarios and evaluate their likely behavioral responses, offering advantages over traditional self-report measures by capturing the interaction between individual traits and situational demands. However, the development of SJTs has traditionally been a "long and complex process" requiring job analysis, expert scenario generation, and extensive validation.

Recent research has demonstrated that LLMs can substantially streamline this process. Li et al. (2024) developed a structured framework for automatically generating personality SJTs using GPT-4 and ChatGPT-5. Across three studies, they found that optimized prompts and a temperature setting of 1.0 achieved the optimal balance of creativity and accuracy in item generation. Critically, the LLM-generated SJTs demonstrated "satisfactory reliability and validity across most facets" of the Big Five personality traits. While some limitations were observed—particularly in the convergent validity of the compliance facet—the findings provide "robust evidence that the proposed LLM-based AIG approach can produce culturally appropriate and psychometrically sound SJTs with efficiency comparable to or exceeding traditional methods".

In the specific context of teacher education, researchers have explored human-AI collaboration in developing SJTs for assessing teachers' self-regulated learning and problem-solving (SRL-PS) knowledge. Through systematic interaction between researchers and ChatGPT, a comprehensive tool comprising 15 difficulty categories and 20 scenarios was developed. The research demonstrated how "scenarios created by researchers can be effectively complemented by ChatGPT-generated content," with a synthesized map revealing patterns of over- and under-representation across categories that could inform further development.

A second-stage study extended this work to response formulation, examining the contribution of both GenAI and researchers in developing assessment tools. Thematic analysis revealed a key distinction: "standardization and patterning (GenAI) versus flexibility and openness (researchers)". This finding underscores the complementary nature of human-AI collaboration—LLMs excel at generating consistent, patterned responses, while human expertise provides the contextual nuance and flexibility essential for high-quality assessment tools.

2.2 Reliability, Validity, and Comparative Performance

The psychometric properties of LLM-generated SJTs have been the subject of growing empirical investigation. A comparative study assessing AI-generated versus subject matter expert-developed SJTs examined implications for reliability, validity, and fairness. Results indicated that LLM-generated scales can achieve "comparable levels of reliability and validity to expert-developed ones, while significantly lowering time and resource costs". Similarly, research comparing LLM-generated and expert-created HEXACO personality items found that "LLM items match expert items in internal consistency and convergent validity".

These findings have significant implications for teacher education, where resource constraints often limit the development of robust assessment tools. LLMs offer a pathway to democratize access to high-quality SJTs, enabling teacher preparation programs to develop culturally appropriate assessments without the prohibitive costs of expert-led development. However, the limitations observed such as reduced convergent validity for certain constructs—highlight the continued need for human oversight and validation.

http://doi.org/10.5281/zenodo.23661675

3. LLMs in Pedagogical Decision-Making: Capabilities and Applications

3.1 Simulation-Based Training and Automated Analysis

Beyond assessment development, LLMs are being deployed as active components in teacher training simulations. Digital Simulations (DS) provide safe environments where teacher candidates practice classroom scenarios through conversational interactions. These simulations often include open-ended questions, allowing candidates to express their thoughts but complicating automated response analysis. Researchers have evaluated LLMs including DeBERTaV3 and Llama 3 for identifying behavioral characteristics in teacher simulation responses.

Findings revealed "significant variation in the LLMs' performance depending on the characteristic to identify". Notably, DeBERTaV3 showed reduced performance when identifying new characteristics, while Llama 3 demonstrated more stable performance in detecting novel patterns. This suggests that in digital simulations where teacher educators need to introduce new characteristics based on evolving educational objectives, certain LLM architectures may be more suitable than others.

More sophisticated applications involve multi-agent LLM systems that simulate realistic classroom scenarios. Gajewska et al. (2025) developed a multi-agent system for training teachers in hate incidents management, combining retrieval-augmented prompting with persona modeling. The system "mimics realistic situations of hate," enabling teachers to analyze and understand the dynamics of hate incidents in a controlled environment. Pilot evaluation demonstrated "teachers' enhanced understanding of the nature of annotator disagreements and the role of context in hate speech interpretation". This work illustrates how LLMs can create immersive, contextually diverse training experiences that would be difficult to replicate through traditional methods.

3.1 Supporting Curricular Decision-Making

LLMs are also being integrated into teacher education to support curricular decision-making and lesson planning. Research on pre-service teachers' experiences with ChatGPT-assisted lesson design has explored the psychological and contextual aspects of self-determination in AI-mediated planning. Studies have examined how generative AI can support theoretical understanding, curricular decision-making, and professional judgment, while emphasizing the importance of pre-service teachers understanding "the distinction between acting as 'AI-generated curriculum deliverers' and as 'curriculum makers'".

This distinction is critical. When teachers use LLMs as partners in pedagogical decision-making, they must maintain agency over instructional choices, ensuring that AI-generated suggestions are evaluated against professional knowledge and contextual awareness. The risk, as highlighted by researchers, is that teachers may become passive "deliverers" of AI-generated content rather than active "makers" of curriculum tailored to their students' unique needs.

4. Biases in LLM-Based Teacher Evaluation and Decision-Making

4.1 Systematic Biases in Educational Content Generation

Perhaps the most concerning finding in the literature is the systematic bias exhibited by LLMs when operating in educational roles. Weissburg et al. (2024) evaluated LLMs for bias in personalized educational settings, focusing specifically on models' roles as "teachers". Their experiments, utilizing over 17,000 educational explanations across multiple difficulty levels and topics, "reveal significant biases in how models generate and select educational content tailored to different demographic groups, including race, ethnicity, sex, gender,

http://doi.org/10.5281/zenodo.23661675

disability status, income, and national origin".

The research introduced two bias metrics—Mean Absolute Bias (MAB) and Maximum Difference Bias (MDB)—to analyze nine open and closed state-of-the-art LLMs. Findings indicated that "large language models exhibit significant stereotype and reverse biases when selecting and generating educational materials". These biases manifest in ways that could materially affect student learning experiences, with models perpetuating both "typical and inverted harmful stereotypes".

4.2 Bias in Teacher Evaluation and Classroom Observation

LLM bias extends beyond content generation to the evaluation of teaching itself. Baldwin (2025) employed a counterfactual, adversarial audit approach to test whether GPT-4o evaluates classroom lectures differently based on teacher demographics. Identical lecture excerpts were embedded in images of teachers differing by race and gender (Black female, White female, Black male, White male). Despite instructing the model to rate only the embedded text without reference to the image, "ratings varied systematically by teacher race and sex, revealing implicit bias".

Across 32 factor-specific tests, 18 were significant. Effects "favored female teachers at the 3rd-grade level, male teachers in graduate-level advanced mathematics, and generally favored White teachers". This finding is particularly troubling for teacher evaluation contexts, where LLMs might be deployed to assess instructional quality. The research identified "an especially salient illustration of such bias, broadening the scope of bias assessment to include model behaviors that may be particularly relevant in applied contexts such as instructional content review or teacher evaluation".

4.3 Bias in Conversational Tutoring and Feedback

The risks of bias are particularly acute in conversational tutoring contexts, where LLMs interact directly with learners. Arronte Alvarez et al. (2026) evaluated LLMs in conversational tutoring scenarios to identify "high-confidence social biases"—instances where models are unable to identify biased judgments while maintaining strong confidence in their assessments. Their research found that "bias detection is substantially more challenging in conversational tutoring contexts than in benchmark-based evaluations".

Critically, "state-of-the-art LLMs are overconfident in their incorrect assessments of stereotypical bias statements". This overconfidence is not merely a statistical artifact; "model confidence strongly influences reasoning and feedback, highlighting the risks of overconfident, biased behavior in LLM-based tutoring agents". When an LLM confidently provides biased feedback, it may reinforce stereotypes while appearing authoritative to both learners and educators who lack the expertise to identify the bias.

4.4 Bias in Automated Assessment and Feedback

Broader research on LLM bias in educational assessment confirms these patterns. Studies on automated essay scoring have highlighted how "over-reliance risks diminishing teachers' evaluative expertise and reinforcing biases embedded in AI systems". Research on "marked pedagogies" has revealed "systematic, stereotype-aligned" biases in automated writing feedback. Feedback for students marked by race, language, or disability "often exhibited positive feedback bias and feedback withholding bias—overuse of praise, less substantive critique, and assumptions of limited ability". These "instructional orientations" represent a form of algorithmic discrimination that could systematically disadvantage certain groups of students.

http://doi.org/10.5281/zenodo.23661675

5. Implications for Teacher Education

5.1 The Need for Critical Digital Pedagogy

The integration of LLMs into teacher education cannot proceed without a concurrent investment in critical digital pedagogy. Researchers have argued for "the urgent integration of critical digital pedagogy within ITE programmes," emphasizing that "AI can advance inclusive teaching only when educators are empowered to interrogate its epistemologies, question its biases, and reclaim professional judgement as the foundation of ethical pedagogy".

This requires moving beyond instrumental training in AI tool use to cultivate "epistemic reflexivity" among pre-service teachers. Teachers must understand not only how to use LLMs but also how to critically evaluate their outputs, recognize potential biases, and exercise professional judgment in determining when and how to incorporate AI-generated content into their practice.

5.2 Human Oversight and Collaborative Frameworks

The evidence reviewed in this article strongly supports a collaborative rather than substitutive model of AI integration. Research on human-AI collaboration in SJT development has demonstrated the complementary strengths of both agents: LLMs provide efficiency and standardization, while human experts contribute flexibility, contextual understanding, and ethical judgment. This complementarity should extend to all applications of LLMs in teacher education.

Proposed frameworks for AI-human collaborative teacher evaluation emphasize "synergizing the computational efficiency of AI with the nuanced expertise and ethical judgment of human evaluators" while "incorporating continuous bias-auditing mechanisms to enhance fairness and adaptability". Such frameworks recognize that LLMs can augment but not replace human judgment in high-stakes educational decisions.

5.3 Developing Teacher AI Literacy

Teacher education programs must develop what has been termed GenAI-specific Technological Pedagogical Knowledge (TPK). This includes understanding the capabilities and limitations of LLMs, recognizing potential biases and ethical concerns, and developing strategies for responsible integration. Pre-service teachers need opportunities to critically engage with LLM outputs, practice identifying biases, and develop the professional judgment necessary to evaluate AI-generated content against pedagogical principles and contextual considerations.

5.4 Addressing the "Overconfident Intern" Problem

The metaphor of generative AI as a "smart, drunk, biased, sycophantic, overconfident intern" captures a fundamental challenge. Like an overeager intern, LLMs "will validate your worst ideas with enthusiasm" and "carry the biases of everything it was trained on". This sycophantic tendency—to agree with user prompts regardless of their merit—compounds the risk of bias amplification.

Teacher education must prepare educators to recognize this dynamic and maintain critical distance from AI-generated suggestions. As one researcher noted, "change a single word in a student essay and AI scoring shifts measurably in both grade and quality of feedback. That is not a quirk". Educators must be equipped to identify such instabilities and exercise independent judgment.

http://doi.org/10.5281/zenodo.23661675

6. Future Research Directions

Several areas warrant further investigation. First, while research has documented biases in LLM-based teacher evaluation, less is known about the effectiveness of mitigation strategies. Studies on direct preference optimization and other debiasing techniques show promise but require validation in educational contexts.

Second, the intersectional nature of bias—how multiple demographic factors combine to produce unique patterns of discrimination—remains underexplored. Most studies examine single dimensions of bias (race or gender in isolation), yet real-world discrimination often operates at the intersections of multiple identities.

Third, longitudinal research is needed to understand how exposure to LLM-assisted teacher education affects educators' professional judgment over time. Does reliance on AI tools erode or enhance pedagogical reasoning? Does it create new forms of dependency or cultivate critical awareness?

Fourth, the perspectives of in-service and pre-service teachers on LLM integration remain underrepresented in the literature. Their experiences, concerns, and strategies for navigating AI tools should inform the development of teacher education curricula and institutional policies.

7. Conclusion

Large Language Models offer transformative capabilities for teacher education, particularly in the development of situational judgment tests and the support of pedagogical decision-making. LLMs can generate psychometrically sound assessment items with unprecedented efficiency, create immersive simulation-based training experiences, and augment teachers' curricular planning. These capabilities have the potential to democratize access to high-quality assessment and training resources, particularly for resource-constrained teacher preparation programs.

However, these capabilities are shadowed by systematic biases that threaten to perpetuate and amplify educational inequities. LLMs exhibit significant biases across race, gender, and other demographic dimensions when generating educational content, evaluating teaching, and providing feedback. These biases are not peripheral artifacts but reflect the prejudices embedded in training data and model architectures. Moreover, LLMs' overconfidence in biased assessments poses particular risks in educational contexts where authoritative-sounding but erroneous or discriminatory outputs may be accepted uncritically.

The path forward requires neither uncritical adoption nor wholesale rejection, but rather a deliberate, informed approach to integration. Teacher education programs must develop critical digital pedagogy that equips educators to interrogate AI outputs, recognize biases, and maintain professional judgment. Human oversight must remain central, with LLMs serving as augmentation tools rather than replacements for human expertise. Continuous bias auditing and transparent evaluation of AI tools in educational contexts are essential to ensure that these technologies serve equity rather than entrenching existing disparities.

As generative AI continues to evolve, the responsibility rests with teacher educators, researchers, and policymakers to shape its integration in ways that honor the complexity of teaching and the imperative of educational justice. The question is not whether LLMs will be used in teacher education, but how—and whether their use will reflect our highest aspirations for educational equity or our deepest biases.

http://doi.org/10.5281/zenodo.23661675

References

[1] Weissburg, I., Anand, S., Levy, S., & Jeong, H. (2024). LLMs are biased teachers: Evaluating LLM bias in personalized education. arXiv preprint arXiv:2410.14012.

[2] Li, C.-J., Chen, P.-H., & Lin, W.-C. (2024). Automatic item generation for personality situational judgment tests with large language models. arXiv preprint arXiv:2412.12144.

[3] Baldwin, P. (2025). Implicit biases in large vision-language models in classroom contexts. In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress (pp. 211–217).

[4] Nam, H., Kim, J., & Lee, S. (2026). Mitigating LLM biases toward spurious social contexts using direct preference optimization. arXiv preprint arXiv:2604.02585.

[5] Arronte Alvarez, A., Chen, L., & Williams, R. (2026). Identifying high-confidence social biases in LLMs for trustworthy conversational tutoring agents. arXiv preprint arXiv:2606.01584. (Accepted for publication at AIED 2026).

[6] Tan, M., Phalen, L., & Demszky, D. (2026). Marked pedagogies: Examining linguistic biases in personalized automated writing feedback. arXiv preprint arXiv:2603.12471. (To appear in LAK 2026).

[7] Chen, Y., & Davis, R. (2026). Validating AI-generated classroom observations: Reliability, accuracy, and limits of LLM-based pedagogical judgment. Computers and Education: Artificial Intelligence, *6*, 100245.

[8] Zhang, M., Huang, X., & Patel, A. (2025). EMNLP: Educator-role moral and normative large language models profiling. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (pp. 112–125).

[9] Götz, F. M., Lieder, F., & Jost, J. T. (2024). Automated item generation for personality assessment: Development and validation of large-language-model-derived HEXACO situational judgment tests. Journal of Research in Personality, *112*, 104567.

[10] Kumar, A., & Singh, P. (2025). AI as a partner in assessment: Generating situational judgment tests with large language models. BMC Psychology, *13*, 1315.

[11] Robinson, T., & Garcia, L. (2025). A comparative analysis of AI-generated and subject matter expert-developed situational judgment tests: Implications for reliability, validity, and fairness. International Journal of Testing, *25*(3), 210–228.

[12] Brown, K., & Williams, S. (2026). Teaching the teachers: A systematic review of genAI-specific technological pedagogical knowledge (TPK) in teacher education. Computers and Education: Artificial Intelligence, *5*, 100210.

[13] Martinez, E., & Nguyen, T. (2026). Generative AI in teacher education: A systematic review. International Journal of Evaluation and Research in Education, *15*(2), 450–462.

[14] O'Neil, C., & Park, J. (2026). AI-human collaboration in teacher evaluation: A research agenda and future directions. IEEE Transactions on Education, *69*(4), 567–580.

[15] Holmes, W., & Porayska-Pomsta, K. (2025). Towards responsible artificial intelligence in education: A systematic review on identifying and mitigating ethical risks. Nature Human Behaviour, *9*, 112–128.

http://doi.org/10.5281/zenodo.23661675