

WIP: Scaffolding Graduate Simulation Education Through Structured LLM Integration and Metacognition: A Preliminary Investigation

Dr. Vishnunarayan Girishan Prabhu, University of Central Florida

Dr. Vishnu Prabhu is an Assistant Professor in the School of Modeling and Simulation at the University of Central Florida. His research focuses on developing digital twins, VR/AR applications, and AI-driven simulations to improve decision-making in healthcare and complex systems. He integrates agent-based modeling, reinforcement learning, and biosensor feedback to create adaptive immersive experiences for clinical training, patient care, and system optimization. Dr. Prabhu has led multiple funded projects, published peer-reviewed research, and mentors students in advanced modeling, simulation, and extended reality technologies.

Roger Azevedo, University of Central Florida

Bulent Soykan

Dr. Ghaith Rabadi, University of Central Florida

WIP: Scaffolding Graduate Simulation Education Through Structured LLM Integration and Metacognition: A Preliminary Investigation

ABSTRACT:

Graduate education in Modeling and Simulation (M&S) faces the challenge of integrating Generative AI (GenAI) to preserve learning objectives while supporting skill development. However, systematic frameworks for pedagogical integration remain underdeveloped. This work-in-progress reports on structured LLM scaffolding in a graduate M&S course using instructor-designed prompts and assignment-embedded constraints. We implemented a within-subjects design where 15 graduate students in an M&S program completed two simulation modeling tasks: (i) one with unrestricted AI use and (ii) one with four structured scaffolding prompts guiding paradigm selection, scoping, validation, and critique. Data included pre- and post-surveys measuring confidence and understanding, correctness of paradigm selection, comparative preferences, qualitative reflections, and AI interaction transcripts. Findings reveal a paradoxical scaffolding effect where students reported higher confidence in work quality with scaffolding (+0.47 on a 10-point scale), but lower confidence in understanding their modeling decisions (-0.40) and a reduced sense that AI helped them learn (-0.20). Different students failed under different conditions with no overlap: three corrected errors with scaffolding, while three made new errors only under scaffolding. Transcript analysis identified four metacognitive patterns: (i) resistance to needed scaffolding despite correction; (ii) disruption of expert strategies, leading to new errors; (iii) false confidence in incorrect scaffolded answers; and (iv) accurate self-assessment of scaffolding benefits and risks. Results suggest that effective AI integration requires explicit teaching of metacognitive awareness, monitoring, regulation, and evaluation, alongside AI literacy skills. Faculty should use adaptive scaffolding that is responsive to demonstrated competence, with explicit metacognitive instruction on recognizing knowledge gaps and critically evaluating AI output.

INTRODUCTION

The rapid adoption of Generative AI (GenAI) tools across higher education has fundamentally altered the landscape of learning [1]. Large Language Models (LLMs), a type of GenAI, have become increasingly popular among both teachers and students across educational contexts [2]. However, this widespread adoption presents faculty with a critical challenge: how to ensure that students learn core concepts and skills rather than copying LLM outputs without understanding them. Traditional binary approaches, either banning the use of AI entirely or allowing free use, miss the pedagogical opportunity to strategically integrate GenAI tools in ways that enhance rather than hinder learning. What is needed are frameworks for faculty to help students responsibly adopt AI use while maintaining academic rigor and integrity [3], [4], [5], [6].

Prior research on AI scaffolding in education has examined contexts such as writing assistance, programming support, and mathematics tutoring, where tasks often involve well-defined procedures and clear evaluation criteria. In contrast, Modeling and Simulation (M&S) education provides a particularly valuable context for investigating AI scaffolding effects because it requires integrating domain knowledge with methodological and analytical skills and expertise across diverse application areas (e.g., K12, healthcare, military, transportation), where students make justified conceptual decisions rather than follow prescribed algorithms. Within M&S, simulation modeling demands these conceptual decision-making skills throughout the modeling process. Fundamental conceptual skills include selecting an appropriate modeling paradigm,

which determines how system behavior is represented and simulated. The three primary paradigms are Agent-Based Modeling (ABM), where autonomous entities interact according to individual rules; Discrete-Event Simulation (DES), where systems evolve through sequences of discrete events; and System Dynamics (SD), where systems are represented by stocks, flows, and feedback loops. Complex problems may require hybrid approaches that combine multiple paradigms. Beyond paradigm selection, students must identify system parameters, define assumptions, and establish validation metrics and performance KPIs. These decisions involve recognizing problem characteristics, evaluating modeling trade-offs, and justifying modeling choices, making it both amenable to LLM support and vulnerable to AI over-reliance [3], [7], [8]. Here, students may accept AI-generated recommendations without critical evaluation, leading to a surface-level understanding and reduced metacognitive engagement in their decision-making [1]. The critical question for faculty, then, becomes: how do we scaffold LLM use to preserve learning and support conceptual understanding and reasoning, while being cognizant that students typically lack the metacognitive skills to make such accurate or optimal decisions?

This work-in-progress paper investigates structured LLM scaffolding through instructor-designed prompts embedded in the course. We focus specifically on conceptual modeling tasks, specifically paradigm selection and model design, rather than coding. Using a within-subjects design, we compare student performance and metacognitive awareness when using unrestricted AI versus scaffolded AI with structured prompts. Our study addresses three research questions: *RQ1: How does structured scaffolding impact the correctness of paradigm selection? RQ2: How does it impact students' perceived confidence in their work quality and understanding? RQ3: Do students accurately recognize when scaffolding helps versus deters their learning processes?*

LITERATURE REVIEW

Students using AI tools for complex conceptual tasks without clear strategies often experience "prompt paralysis," uncertain how to engage productively with LLMs. Scaffolding, the provision of temporary support structures that fade as learner competence develops, addresses this challenge [10]. In the context of AI tool use, structured prompts guide systematic thinking by breaking problems into steps, prompting comparison of alternatives, and encouraging reasoning. Cognitive load theory provides a theoretical framework for understanding why such scaffolding works: by reducing extraneous cognitive load, instructional design frees working memory for schema development [9]. However, scaffolding is not always beneficial. The expertise reversal effect reveals that instructional support helpful to novices can impede experts by disrupting their advanced schemas [11]. Further, over-scaffolding may prevent students from developing independent strategic thinking and metacognitive skills that define expertise, suggesting that effective AI scaffolding must be adaptive rather than one-size-fits-all [12].

The effectiveness of scaffolding depends on metacognitive awareness, the ability to monitor and regulate one's own cognitive processes, which is central to self-regulated learning, where learners plan, monitor, and evaluate their strategies and outcomes [12]. Prior studies demonstrate that experts and novices differ not only in domain knowledge but also in metacognitive calibration, in which experts accurately judge when they need help or additional information, while novices often exhibit poor awareness of their knowledge gaps [13]. Research with intelligent tutoring systems demonstrates that scaffolding metacognitive processes significantly enhances learning outcomes [12], [14]. During LLM use, students face a novel metacognitive challenge: deciding when to accept versus critically evaluate AI output. Without accurate metacognitive awareness,

students risk what has been termed "metacognitive laziness": overreliance on AI outputs without engaging the critical metacognitive skills and the cognitive effort necessary for learning [8].

Despite growing interest in AI-augmented learning, most research examines aggregate effects rather than individual differences, i.e., when does scaffolding help versus hurt, and for whom? Research has not examined whether students accurately recognize when they need structured guidance versus free exploration. From an M&S perspective, no prior work has investigated how LLM scaffolding impacts conceptual modeling skills, such as paradigm selection, which requires integrating problem understanding with methodological knowledge. This work-in-progress addresses these gaps by examining both learning outcomes (paradigm selection correctness) and metacognitive awareness (self-assessment accuracy, accurate selection of tool preferences, and recognition of being supported or constrained by scaffolding), providing preliminary evidence to inform faculty decisions about when and how to scaffold LLM use in graduate M&S education.

METHODS

Participants and Context

This study was conducted in a graduate M&S course at University of Central Florida during Fall 2025. Fifteen graduate students completed the study with no dropouts; most were experienced (67% completed 4+ M&S courses) and frequent AI users (60% used AI daily/weekly).

Study Design and Scenarios

We employed a within-subjects design in which all participants completed two conceptual modeling tasks under different conditions. Scenario A (vaccine distribution) allowed unrestricted AI use, while Scenario B (wildfire evacuation) provided four instructor-designed scaffolding prompts that systematically guided paradigm selection, model scoping, parameter identification, and validation planning. Both scenarios were designed to favor hybrid approaches combining ABM and DES. Specifically, agents represent heterogeneous entities with individual decision-making behaviors, while DES captures time-dependent event processes. Each scenario had an approximate expected task completion time of 20 minutes.

Data Collection and Analysis

We collected data from multiple sources at different time points. Before the scenarios, students completed pre-surveys assessing their AI experience and baseline confidence. For each scenario, students submitted their conceptual model designs and AI chat transcripts, then completed post-scenario self-assessments rating their confidence in work quality, understanding of the system, understanding of modeling decisions, and the extent AI helped them learn (all on 1-10 scales). After completing both scenarios, students provided comparative preferences between the two approaches and qualitative reflections on their experiences. Paradigm selections were expert-coded as correct (hybrid approach matching problem characteristics), partially correct (single paradigm with acknowledged limitations), or incorrect (paradigm misaligned with problem requirements). We analyzed self-assessment metrics descriptively and conducted detailed transcript analysis for students who made errors in either condition.

RESULTS

RQ1: Impact of Scaffolding on Paradigm Selection Correctness

We first examined whether scaffolding improved students' ability to select appropriate modeling paradigms. Expert review showed that 12 of 15 students selected correct paradigms in both

conditions, with non-overlapping errors: 3 students (20%) who selected incorrect paradigms under unrestricted AI corrected their choices with scaffolding, while 3 different students (20%) who correctly identified paradigms under unrestricted AI made errors with scaffolding. None made errors in both conditions, indicating that scaffolding does not uniformly improve or impair performance. Instead, it appears to enhance accuracy for some while constraining others.

RQ2: Impact on Perceived Confidence and Understanding

We next analyzed how scaffolding affected students' self-assessed confidence in their work quality and understanding of their modeling decisions. Analysis of post-scenario surveys revealed a paradoxical pattern where students reported higher confidence in work quality under scaffolding, yet lower confidence in their understanding and a sense that AI helped them learn (Table 1). Specifically, scaffolded prompts improved confidence in model quality by 0.47 points, suggesting students perceived scaffolded outputs as more robust. However, understanding of modeling decisions decreased by 0.40 points, and perception that "AI helped me learn concepts" decreased by 0.20 points. This inverse relationship between improved artifact confidence and lower learning confidence suggests that scaffolding may create a false sense of accomplishment, leading students to trust their outputs more but feel less ownership of the reasoning behind them.

Table 1: Self-Assessment by scenario.

Metric	Scenario A (Unrestricted AI)	Scenario B (Scaffolded AI)	Difference (B - A)
Confidence in the quality of the model	6.93	7.40	+0.47
Understanding of the system	6.87	6.80	-0.07
Understanding of modeling decisions	7.60	7.20	-0.40
AI helped me learn/understand concepts	6.87	6.67	-0.20

RQ3: Metacognitive Awareness of Scaffolding Effects

Finally, we examined whether students correctly recognized when scaffolding supported or constrained their learning. When asked about preferences for AI scaffolding, 53% of students wanted a "50/50 mix depending on topic complexity," with no students wanting scaffolding always or never. This suggests students recognize that scaffolding utility varies by context, even if they cannot predict when they need it. Detailed transcript analysis of the six students who made errors revealed four distinct metacognitive patterns, demonstrating systematic failures in students' recognition of when they needed scaffolding versus when it constrained them.

Pattern 1: The Blind Spot (n=2). Students who made errors under unrestricted AI were corrected by scaffolding, yet rejected its value. They were committed to paradigms before systematic analysis and never questioned whether single approaches were sufficient. Despite scaffolding correcting their errors, they rated unrestricted AI use as more helpful for their understanding, and scaffolding was "never" valuable, indicating blind spots in metacognitive awareness.

Pattern 2: The Constrained Expert (n=1). One student demonstrated sophisticated research-first strategies in unrestricted conditions, gathering empirical case studies before making paradigm decisions, and correctly selected a hybrid approach. Under scaffolding, this student accepted a pure ABM recommendation, with understanding ratings dropping from 7/10 to 1/10. Qualitative feedback indicated scaffolding felt like "following commands blindly," demonstrating that structure can disrupt effective expert workflows.

Pattern 3: Rare Insight (n=1). One student demonstrated metacognitive awareness, recognizing the benefits and risks of scaffolding. This student acknowledged that structured prompts ensured coverage but also created temptation to "transpose answers over from AI output" rather than engage deeply. They recommended adaptive scaffolding for "specific components" rather than "entire systems," demonstrating awareness of when structure helps versus overwhelms.

Pattern 4: False Confidence (n=2). Students who answered correctly under unrestricted conditions but incorrectly under scaffolding rated their wrong scaffolded work higher than their correct unrestricted answers. One student rated correct work 6/10 but incorrect scaffolded work 8/10, stating a preference for "always" using structured prompts. This reveals how detailed, well-organized AI responses can create a false understanding even when conclusions are incorrect.

These patterns demonstrate that metacognitive miscalibration occurs in both directions, where students who need scaffolding resist it, students who do not need it embrace it, and effective strategies get disrupted by structure [15], [16]. Only one student demonstrated accurate awareness, suggesting that metacognitive calibration must be explicitly taught [17], [18].

DISCUSSION, LIMITATIONS AND CONCLUSIONS

Our findings reveal a scaffolding paradox: structured prompts increase confidence in work quality while decreasing students' sense of learning and understanding. This suggests "confident outsourcing," where detailed AI responses signal quality while reducing ownership of the reasoning. However, this aggregate pattern masks critical individual variation in how scaffolding affects different students. More critically, the complete lack of overlap between students who failed under unrestricted versus scaffolded conditions demonstrates that one-size-fits-all approaches fundamentally misunderstand how students engage with AI tools. Some students need structure to overcome pre-commitment biases (Blind Spot), while others have effective strategies that scaffolding disrupts (Constrained Expert). The prevalence of calibration challenges the assumption that metacognitive awareness develops naturally from AI experience. These findings suggest that faculty must teach metacognitive knowledge and skills, particularly metacognitive calibration, explicitly alongside AI tool use. Students need frameworks for recognizing when they need structure versus when it constrains them, such as metacognitive decisions in which students predict their confidence before and after AI interaction, compare predictions to performance, and reflect on calibration accuracy. Scaffolding should be adaptive and targeted for foundational decisions that structure future work, rather than comprehensive.

This paper has several limitations that we are addressing in our continued investigation. The small sample size (n=15) and non-counterbalanced design limit generalizability, though the within-subjects approach and qualitative data provide meaningful preliminary insights into individual differences. Our next phase will expand the sample to test findings across multiple course sections and implement counterbalancing to control scenario effects. We are currently designing adaptive scaffolding mechanisms that respond to demonstrated competence rather than applying a uniform structure, informed by the distinct patterns identified here. Additionally, we plan to develop and test explicit metacognitive training interventions that teach students to recognize when they need structure versus when it constrains them, with longitudinal assessment of whether calibration accuracy improves over a semester. Finally, we are exploring real-time feedback mechanisms that could help students monitor their AI engagement during tasks rather than only in retrospective reflection.

References

- [1] M. Giannakos *et al.*, “The promise and challenges of generative AI in education,” *Behaviour and Information Technology*, vol. 44, no. 11, pp. 2518–2544, Jul. 2025, doi: 10.1080/0144929X.2024.2394886.
- [2] D. L. Dinsmore and L. K. Fryer, “What does current genAI actually mean for student learning?,” *Learn. Individ. Differ.*, vol. 125, p. 102834, Jan. 2026, doi: 10.1016/J.LINDIF.2025.102834.
- [3] V. G. Prabhu, B. Soykan, R. Azevedo, S. Mondesire, and G. Rabadi, “Teaching Simulation with AI: A Faculty-First Framework for Integrating LLMs into Coursework,” Florida A.I. Learning Consortium (FALCON). [Online]. Available: <https://www.fl-falcon.org/teaching-simulation-with-ai-a-faculty-first-framework-for-integrating-llms-into-coursework/#respond>
- [4] J. McGinty and K. Murray-Johnson, “Ethical Frameworks of Artificial Intelligence for Faculty: Upholding Academic Integrity and Authenticity,” *New Directions for Adult and Continuing Education*, vol. 2025, no. 188, pp. 15–23, Dec. 2025, doi: 10.1002/ACE.70012.
- [5] H. Albasry, E. Carmona-Cejudo, A. Rauf, and D. Chen, “A systematically derived AI-based framework for student-centered learning in higher education,” *Social Sciences & Humanities Open*, vol. 12, p. 102085, Jan. 2025, doi: 10.1016/J.SSAHO.2025.102085.
- [6] A. Tolk, P. Barry, M. L. Loper, G. Rabadi, W. T. Scherer, and L. Yilmaz, “Chances and Challenges of ChatGPT and Similar Models for Education in M&S,” in *Proceedings - Winter Simulation Conference*, C. G. Corlu, S. R. Hunter, H. H. Lam, B. S. Onggo, J. Shortle, and B. Biller, Eds., San Antonio: Institute of Electrical and Electronics Engineers Inc., Dec. 2023, pp. 3332–3346. doi: 10.1109/WSC60868.2023.10408241.
- [7] A. Akhavan and M. S. Jalali, “Generative AI and simulation modeling: how should you (not) use large language models like ChatGPT,” *Syst. Dyn. Rev.*, vol. 40, no. 3, p. e1773, Jul. 2024, doi: 10.1002/SDR.1773;ISSUE:ISSUE:DOI.
- [8] Y. Fan *et al.*, “Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance,” *British Journal of Educational Technology*, vol. 56, no. 2, pp. 489–530, Mar. 2025, doi: 10.1111/BJET.13544;PAGE:STRING:ARTICLE/CHAPTER.
- [9] J. Sweller, “Cognitive Load Theory,” *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 55, pp. 37–76, Jan. 2011, doi: 10.1016/B978-0-12-387691-1.00002-8.

- [10] C. C. A. van Nooijen *et al.*, “A Cognitive Load Theory Approach to Understanding Expert Scaffolding of Visual Problem-Solving Tasks: A Scoping Review,” *Educational Psychology Review* 2024 36:1, vol. 36, no. 1, pp. 12-, Jan. 2024, doi: 10.1007/S10648-024-09848-3.
- [11] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, “The Expertise Reversal Effect,” *Educ. Psychol.*, vol. 38, no. 1, pp. 23–31, 2003, doi: 10.1207/S15326985EP3801_4.
- [12] R. Azevedo *et al.*, “Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning with an Intelligent Tutoring System,” *Front. Psychol.*, vol. 13, p. 813632, Jun. 2022, doi: 10.3389/FPSYG.2022.813632/FULL.
- [13] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams, “The Cambridge Handbook of Expertise and Expert Performance: Second Edition,” *The Cambridge Handbook of Expertise and Expert Performance: Second Edition*, pp. 1–970, Jan. 2018, doi: 10.1017/9781316480748.
- [14] R. Azevedo and M. Wiedbusch, “Theories of metacognition and pedagogy applied to AIED systems in: Handbook of Artificial Intelligence in Education,” in *Sociology, Social Policy and Education*, vol. 1, Cheltenham: Edward Elgar Publishing, 2023, ch. 4, pp. 45–67. Accessed: Jan. 15, 2026. [Online]. Available: <https://www.elgaronline.com/edcollchap/book/9781800375413/book-part-9781800375413-13.xml>
- [15] S. M. Fleming and H. C. Lau, “How to measure metacognition,” *Front. Hum. Neurosci.*, vol. 8, no. JULY, p. 82285, Jul. 2014, doi: 10.3389/FNHUM.2014.00443/TEXT.
- [16] D. J. Hacker, L. Bol, and M. C. Keener, “Metacognition in Education: A Focus on Calibration,” in *Handbook of Metamemory and Memory*, First., J. Dunlosky and R. A. Bjork, Eds., New York: Psychology Press, 2008, pp. 429–455. doi: 10.4324/9780203805503-25.
- [17] S. Tobler and M. Kapur, “Metacognitive Calibration: A Methodological Expansion and Empirical Application,” in *Proceedings of International Conference of the Learning Sciences, ICLS*, P. Blikstein, J. Van Aalst, R. Kizito, and K. Brennan, Eds., Montreal: International Society of the Learning Sciences, 2023, pp. 3–10. doi: 10.22318/ICLS2023.352932.
- [18] J. Dunlosky and K. A. Rawson, *The cambridge handbook of cognition and education*. Cambridge University Press, 2019. doi: 10.1017/9781108235631.