2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #50514

AI as a Constrained Guide: Measuring Knowledge Transfer in a Graduate Engineering Course

Dr. Edgar Lobaton, North Carolina State University at Raleigh

Edgar J. Lobaton is a Professor in the Department of Electrical and Computer Engineering (ECE) at North Carolina State University (NCSU). He joined the department in 2011. Lobaton earned his B.S. in Mathematics and Electrical engineering from Seattle University in 2004. He completed his Ph.D. in Electrical Engineering and Computer Sciences from the University of California, Berkeley in 2009. Lobaton was engaged in research at Alcatel-Lucent Bell Labs in 2005 and 2009. He was awarded the NSF CAREER Award in 2016. He was also awarded the 2009 Computer Innovation Fellows post-doctoral fellowship and conducted research in the Department of Computer Science at the University of North Carolina (UNC) at Chapel Hill from 2009 until 2011. In 2024, he received the University Faculty Scholars and the Outstanding Teacher Awards from NC State. His research focuses on the integration of AI, and physical and probabilistic modeling applied to cyber-physical systems in areas such as wearable health monitoring, rehabilitation robotics, agriculture and biological imaging.

Mayur Sanap, North Carolina State University at Raleigh

Mayur Sanap is a Ph.D. student in Electrical Engineering at North Carolina State University. His research interests include edge AI, AI for healthcare, and embedded systems. In this paper, he contributed by developing the AI metrics and conducting a literature review of prior work.

Ms. Ishita Pai Raikar, North Carolina State University at Raleigh

Ishita Pai Raikar is a Computer Engineering student at North Carolina State University with a minor in Philosophy. Her research focuses on AI ethics, human-robot interaction, and responsible AI integration in educational and social systems. She works with the NeuroComputational Ethics Research Group, where she studies ethical frameworks for socially assistive robotics and the moral dimensions of AI-mediated decision-making. She also leads AI Revolution: an #EthicsQuest Across Disciplines, a panel series that brings together academic, industry, and policy experts to examine the future of human-centered and ethically aligned AI. Her broader interests include AI-enhanced learning, interdisciplinary collaboration, and the societal impacts of emerging technologies.

Dr. Carine Marette, Indiana University-Bloomington

Carine Marette is a doctoral candidate in Instructional System Technology at Indiana University Bloomington's School of Education, specializing in AI-enhanced engineering pedagogy and learning analytics. Her research examines how constrained artificial intelligence environments can function as measurement affordances for observing transfer-relevant behaviors in authentic learning contexts. She developed the Vector Framework for multi-source learning telemetry, integrating behavioral data from AI interaction logs and peer assessment platforms to measure transfer readiness, productive struggle, and metacognitive development. Her work bridges engineering education, cognitive load theory, and educational measurement, with particular focus on developing validity arguments for AI-mediated assessment in project-based learning environments. Carine co-created with Kritik360 and VisibleAI platforms, which provide transparency-oriented authorship tracking and structured peer evaluation systems. Her dissertation investigates adaptive constraint design for differentiated AI scaffolding across learner distributions in graduate engineering courses.

Md Halim Mondol, North Carolina State University at Raleigh

Md Halim Mondol is a PhD student in the Department of Electrical and Computer Engineering at North Carolina State University. His research centers on machine learning, large language models, and generative models, particularly their application to knowledge hub development and cheminformatics, aiming to advance data-driven discovery and intelligent information systems.

AI as a Constrained Guide: Measuring Knowledge Transfer in a Graduate Engineering Course

Abstract

This paper presents a novel pedagogical experiment in a graduate-level engineering course, designed to assess how students integrate newfound knowledge and expertise when utilizing conversational AI as a constrained project assistant. Our primary goal was to move beyond the typical "AI as a tutor" model to establish the AI as an effective "guide," requiring students to generate content from their own knowledge base while using the AI solely for structural critique, redirection, and resource navigation. The core of the study involved a structured, two-phase project proposal exercise. Early in the semester, before formal introduction to core course topics, students drafted an initial proposal for a physics-aware deep learning project, leveraging a conversational AI assistant whose information output was programmatically limited (e.g., restricted to providing only conceptual definitions or structural feedback). A second, identical proposal was drafted at the end of the semester after all technical topics had been covered. We hypothesize that the knowledge gained throughout the semester would significantly improve the quality and technical depth of the final proposal. More critically, we aim to decouple the effect of knowledge gained (understanding concepts) from expertise gained (skilled interaction with the AI tool). The evaluation utilizes a mixed-methods approach. Quantitative metrics are derived from the students' tracked interactions with the conversational AI. Qualitative and self-reported measures are collected via surveys at the end of each assignment, which capture their evolving confidence in using the AI as an assistant rather than a solution provider.

Introduction

The rapid integration of generative AI tools—particularly large language models (LLMs) such as ChatGPT, Claude, and Copilot—into engineering education is reshaping how students learn, practice, and demonstrate competence[1,2]. These systems can support brainstorming, critique, and productivity, offering new opportunities for personalized learning and enhanced problem solving. At the same time, their adoption introduces a critical assessment challenge: determining whether a student's performance reflects genuine mastery or effective AI assistance[3]. In engineering education, this challenge is closely tied to transfer, as engineers are expected to apply foundational knowledge to novel, ill-structured problems across diverse contexts—the very type of competence that AI tools can sometimes replicate superficially[4,5]. Traditional assessments often emphasize final products or isolated skills, yet they do not reliably reveal whether students can generalize learned concepts to new situations without AI scaffolding[6]. This concern is especially salient in graduate engineering programs, where students are expected to develop expertise that supports independent problem solving and reasoning in unfamiliar domains[7].

A second challenge arises from the dual nature of AI-assisted learning. Students may simultaneously develop conceptual knowledge in the domain and AI interaction expertise, which includes the ability to prompt effectively, critique outputs, and iteratively refine work using AI feedback[2]. Current assessment practices frequently conflate these competencies, making it difficult to determine whether improved performance reflects deeper conceptual understanding or increased skill in using the tool. This conflation risks overestimating mastery when strong AI-use skills compensate for weak conceptual understanding, and underestimating learning gains when students improve conceptually but remain inefficient in AI interaction. These risks have implications for instructional decision-making and credential validity, particularly if course outcomes implicitly certify tool-based performance rather than engineering competence[3]. Existing approaches for addressing AI involvement are limited: self-reported measures are susceptible to social desirability bias, and AI-detection tools remain unreliable due to high error rates and instability across contexts[8]. These limitations underscore the need for assessment methods that directly measure transfer-relevant competence rather than inferring learning from AI usage alone.

Recent literature on AI-assisted engineering education highlights several gaps that motivate this work. First, transfer is rarely measured directly; many studies focus on engagement, satisfaction, or single-task performance rather than evaluating whether learning generalizes beyond the immediate activity[9]. Second, few validated instruments exist that can disentangle conceptual knowledge from AI interaction expertise, despite growing recognition that these are distinct competencies. Third, constrained AI environments remain understudied, even though pedagogical constraints that limit AI to guidance rather than solution generation may reduce over-reliance and better support robust learning[10]. Together, these gaps indicate the need for assessment tools capable of evaluating whether students can demonstrate transferable engineering reasoning when AI is present but bounded.

To address these challenges, we implement several rubric-based metrics (including quality, depth, feasibility, and structure) designed to measure transfer readiness in AI-assisted engineering contexts. These metrics are evaluated using a structured, two-phase proposal exercise that captures student performance before and after instruction while also tracking interactions with a constrained conversational AI assistant. Accordingly, we investigate the following research question: How does the quality of graduate engineering project proposals change over the course of a semester when students use a conversational AI system configured as a constrained guide? We further examine whether these metrics correlate with peer evaluation metrics aimed at capturing conceptual learning gains. The use of these tools is explored as part of the evaluation of a graduate course with a diverse student population.

Methods

Course Context

The course examined in this study is the graduate-level *Physics Modeling with Neural Networks* course at NC State. This 1-credit course introduces techniques for modeling physical systems using neural networks, including neural ordinary differential equations (Neural ODEs), physics-informed neural networks (PINNs), deep neural operators, and approaches to uncertainty quantification. The course emphasizes both theoretical foundations and hands-on implementation using state-of-the-art software libraries.

The learning objectives for the course are to:

1. Develop an understanding of core concepts on neural networks for modeling physical systems.
2. Explain the fundamental ideas underlying Neural ODEs, PINNs, and Neural Operators.
3. Demonstrate basic proficiency with the libraries used to implement the techniques discussed in class.

Prerequisites include prior experience selecting, designing, and training neural networks, as well as proficiency in Python programming. The study was conducted during the Fall 2025 semester with an enrollment of 18 students representing several departments, including Electrical and Computer Engineering, Computer Science, Mechanical Engineering, and Materials Science and Engineering.

Course activities included multiple programming assignments and two project proposal submissions—one at the beginning of the semester (pre) and one at the end (post). Both proposals were completed with the support of an AI system configured to act as a constrained guide, and these assignments form the basis of our analysis.

Platform for AI-Guided and Peer Review

The proposal assignments in this study were completed using the Kritik360 and VisibleAI platforms from Kritik[11]. Together, these platforms are designed to support AI-guided learning and structured peer assessment.

Kritik360 is a peer-assessment platform that enables scalable learning while reshaping the instructor's role. Each Kritik360 activity follows a three-stage workflow—create, evaluate, and feedback. Students first submit an original artifact, then anonymously evaluate several peers' submissions using an instructor-defined rubric, and finally reflect on and respond to the feedback they receive. This workflow operationalizes core principles of peer learning by requiring students to interpret criteria, judge quality, and communicate actionable suggestions, thereby strengthening evaluative judgment and feedback literacy. The structure aligns with *formative assessment theory*[12] and *constructivist engineering pedagogy*[13], both of which emphasize iterative cycles of production, critique, and revision. By engaging students in repeated evaluation and justification of their decisions, the platform cultivates competencies central to engineering practice, including effective communication, teamwork, ethical and responsible judgment, and the ability to assess and improve technical work. These activities also contribute to the development of *engineering epistemic cognition*[14], or the ways in which engineers reason about evidence, validity, and design quality.

VisibleAI, a transparency-oriented authorship analysis tool integrated into the creation stage, provides a structured breakdown of how AI contributed to each student submission. It distinguishes among original student content, AI-generated content, AI-revised content, and pasted content. Rather than serving as a detection or policing mechanism, VisibleAI is designed to promote responsible AI use by making AI contributions visible and interpretable to both students and instructors. This approach aligns with emerging calls in AI-in-education research for explainable, human-centered systems that foster appropriate trust, interpretation, and oversight of AI-supported learning processes.

A distinctive feature of Kritik360 is its approach to accountability and scoring reliability in peer assessment. Students receive scores reflecting their performance across the creation, evaluation, and feedback stages. The platform employs a weighted peer-grading mechanism in which evaluations from more accurate reviewers carry greater influence on final scores. This design draws on principles from *measurement theory*[15,16] and *assessment for learning*[12], enabling peer assessment to remain both scalable and psychometrically defensible—an important consideration in large engineering courses. Sofjan et al.[17] demonstrated strong correlations between weighted peer scores and instructor evaluations, providing empirical support for the reliability and scalability of the Kritik360 model. These findings motivate the use of Kritik360 and VisibleAI as a combined infrastructure for fostering reflective, accountable, and AI-aware learning in graduate engineering contexts.

The integrated use of VisibleAI and Kritik360 created a learning environment in which AI functioned as a *constrained guide* rather than a tutor or answer-providing system. In this course, this design supported the study's goal of disentangling students' mastery of domain knowledge from their proficiency with AI tools. It enabled instructors to observe how students increasingly relied on their own conceptual understanding while using AI for structure, clarification, and refinement. This framing aligns with research on *knowledge transfer*, which emphasizes the importance of students applying learned concepts independently rather than outsourcing cognitive work to tools.

Assignment and Data Collection Procedure

The students were provided with the following instructions for the proposal assignments:

> *The objective of this assignment is to develop a project proposal on a topic of your choice that involves the use of physics modeling using neural networks for an engineering application. Make sure to pick an area that you are interested in but that is outside your area of expertise. You are required to use the provided AI assistant as a tool for brainstorming, clarifying concepts, and refining your ideas.*

Your Task

You will prepare a 500-word project proposal (no references needed) that outlines a research project you would like to pursue. Your proposal should be well structured and include the following sections:

- ***Introduction and Motivation:*** *Briefly introduce the engineering problem you aim to solve and explain its significance.*
- ***Problem Statement and Objectives:*** *Clearly define the problem and list the specific objectives of your project.*
- **Methodology:** Describe the neural network architecture you plan to use and the data you will need.
- ***Expected Outcomes and Impact:*** *Discuss the expected results of your project and their potential impact on the field.*

Instructions for Using the AI Assistant

The system tracks the amount of content generated by AI, so do NOT just copy the assignment description in an AI assistant and paste the response. You are encouraged to use an AI assistant throughout the proposal development process. Here are some examples of how you can leverage this tool:

- ***Brainstorming:*** *Use the AI to generate a list of potential project ideas at the intersection of physics, neural networks, and engineering. Make sure to pick a topic area outside your area of expertise. Here are some sample prompts to get you started:*
  - *"What are some emerging applications of physics-based machine learning in biomedical engineering? I'm particularly interested in topics like modeling blood flow, tissue mechanics, or medical imaging."*
  - *"Provide some innovative ideas for using PINNs to address problems in sustainable energy systems. Think about areas like battery performance modeling, wind turbine optimization, or grid stability."*
- ***Refining Your Research Question:*** *Once you have a general idea, work with the AI to narrow down your topic and formulate a clear and concise research question.*
- ***Clarifying Concepts:*** *If you are unsure about a particular concept, ask the AI to explain it in simple terms or provide an analogy.*

The assignment rubric used a 0–4 scale, with criteria and descriptions for the extreme values summarized in Table 1. All assignments were completed within the VisibleAI platform, which provided a constrained, AI-guided environment for proposal development. Grading was conducted through the Kritik360 platform, where each submission was evaluated by three randomly assigned peers. For analysis, individual criterion scores are normalized to a 0–100%

Table 1: Range of Proposal Rubric

Criteria	Level 0	Level 4
Introduction and Motivation	The introduction lacks a clear explanation of the engineering problem or its significance.	The introduction provides a compelling and insightful overview of the engineering problem and its global significance.
Problem Statement and Objectives	No clear problem statement is provided, and the objectives are not defined.	The problem statement is expertly crafted with precise and ambitious objectives.
Methodology	No description of the neural network architecture or data requirements is included.	A detailed and innovative description of the neural network architecture and data requirements is presented.
Expected Outcomes and Impact	The expected outcomes and their impact are not discussed.	Thoroughly articulated outcomes with a deep analysis of their transformative impact.
Writing Clarity and Organization	The proposal is poorly organized and difficult to follow, with numerous grammatical errors.	The proposal is exceptionally clear and well-organized with flawless grammar.
Originality and Creativity	The proposal lacks originality, with little to no evidence of creative thought.	The proposal is highly original, showcasing groundbreaking and imaginative ideas.

scale. The overall assignment score is computed as the unweighted average of the criterion-level scores.

AI Interaction Metrics

To characterize how students engage with a conversational AI under constrained assistance, we define a set of interaction metrics that capture both *learning outcomes* (improvements in proposal quality) and *process signals* (the extent to which students iteratively use the AI as a guide rather than a solution provider). Our overarching goal is to measure *knowledge transfer* from course instruction into the final proposal while disentangling this effect from students' growing *expertise in interacting with the AI tool*. This section motivates candidate metric families based on prior work in intelligent tutoring systems (ITS) and LLM-supported learning, and then introduces the focused subset used in our analysis.

Research in ITS has demonstrated that learning-relevant behaviors can be inferred from structured interaction traces. D'Mello and Graesser[18] modeled affective trajectories during complex learning (e.g., confusion→engagement versus confusion→frustration), showing that turn-taking patterns and clarification sequences can serve as behavioral signatures of productive struggle or disengagement. Aleven et al.[19] similarly operationalized help-seeking behaviors from tutoring logs—such as hint frequency, latency between hints, rapid-fire clicking indicative of frustration, and avoidance patterns—linking these indicators to scaffolding effectiveness and feedback-driven revision. Extending this perspective, D'Mello and Graesser[20] conceptualized confusion as an epistemic emotion arising from cognitive disequilibrium, distinguishing productive confusion that resolves and supports learning from persistent confusion that escalates toward frustration or disengagement.

Recent work on generative AI-assisted learning highlights that both interaction quality and perceived output quality influence learning outcomes through motivational and self-efficacy pathways. Bai and Wang[21] find that higher-quality interactions and outputs from generative AI tools are associated with increased learning motivation and creative self-efficacy, with motivation mediating the relationship between AI output quality and learning gains. In parallel, rubric-based LLM assessment has emerged as a scalable method for evaluating open-ended student work. Agostini et al.[22] replicate prior findings showing that LLMs can reliably score authentic student tasks using instructor-defined rubrics, supporting our use of multi-dimensional rubric scoring for pre/post proposal evaluation and diagnostic analysis.

From this broader set of candidate indicators, we implement a focused subset spanning *outcome*, *process*, and *process–outcome* constructs. These are the metrics considered:

1. **Quality Score (Q).** Q serves as the primary outcome metric, capturing overall proposal quality for each student at both the pre- and post-proposal stages. It provides a normalized measure of quality on a continuous scale from 0 to 1 and forms the basis for computing learning gain. The metric is designed to be rubric-consistent, model-consistent, and independent of student–AI interaction data, enabling it to function as a clean learning-outcome variable. Each assignment was evaluated using a fixed large language model configured as a rubric-based grader. The same model and grading prompt were consistently used to compute both the overall Quality Score (Q) and the individual rubric-based metrics of technical depth, feasibility, and structure/clarity. The grading prompt was based on a graduate-level engineering proposal rubric, from which these dimensions were directly extracted alongside the overall quality score. The learning gain was computed as

$$\Delta Q \;=\; Q_{\text{post}} - Q_{\text{pre}}. \tag{1}$$

   To ensure reproducibility and consistency, all proposals were graded using identical prompts and model parameters, and grading outputs were cached using a hash of the proposal text.

2. **Technical Depth.** Depth measures the extent to which a student's proposal demonstrates substantive engagement with the underlying technical concepts relevant to the project domain. This includes the appropriate use of mathematical formulations, algorithmic reasoning, modeling assumptions, and domain-specific terminology expected at the graduate level. The metric is designed to capture conceptual rigor rather than surface-level description. Depth and Δ Depth are obtained in a similar way how Q is computed.

By isolating depth as a distinct dimension, this metric enables analysis of whether students moved beyond high-level descriptions toward more precise, technically grounded proposals over the course of the semester. Importantly, D reflects learning outcomes attributable to conceptual understanding and is evaluated independently of conversational interaction volume.

3. **Feasibility (Feas).** It evaluates the extent to which a proposed project is practically realizable within the given constraints, including scope, methodological appropriateness, and resource considerations. This metric captures whether the proposed approach is coherent, logically structured, and achievable within a typical graduate-level project timeline.

   This metric enables examination of whether students developed a more realistic and methodologically sound project plan over the semester. Importantly, feasibility reflects students' ability to translate conceptual understanding into an executable proposal and is evaluated independently of the amount of interaction with the conversational AI.

4. **Structure (Struct).** It assess the organizational quality and communicative effectiveness of each proposal. This metric captures whether the document presents a logically ordered narrative, clearly defined sections, and coherent progression from motivation through methodology and expected outcomes.

   This metric reflects students' ability to organize technical content into a clear and persuasive proposal format. Improvements in structure indicate increased proficiency in articulating ideas, structuring arguments, and presenting complex material in a manner appropriate for a graduate-level engineering audience, independent of interaction frequency with the AI assistant.

5. **Total Student Turns (I).** It is defined as the total number of student-generated chat messages (i.e., individual user inputs) during interaction with the conversational AI system. It serves as a coarse measure of interaction intensity rather than learning quality. This metric captures how much the AI was used, independent of proposal content, grading outcomes, or conversational effectiveness.

   By construction, I is non-negative and unbounded, reflecting variability in students' interaction strategies and help-seeking behavior. Importantly, I is not interpreted as a performance metric; rather, it is used as a contextual process variable to distinguish efficient knowledge integration from excessive or unfocused AI reliance.

6. **Efficiency of Iteration (EOI).** The EOI quantifies how effectively students translated interaction with the conversational AI into measurable learning gains. Unlike raw interaction counts, EOI normalizes proposal quality improvement by the volume of student–AI exchanges, thereby distinguishing productive engagement from overuse.

   Let I denote the total number of student turns during AI interaction, and let ΔQ represent the learning gain in proposal quality. EOI is defined as

$$\text{EOI} = \frac{\Delta Q}{1 + \log(1 + I)}. \tag{2}$$

This formulation penalizes excessive interaction while preserving sensitivity to meaningful improvement, ensuring that higher EOI values correspond to efficient, goal-directed use of the AI assistant. Importantly, EOI is a process-aware metric: low values may reflect inefficient prompting strategies or exploratory overuse rather than lack of ability. As such, EOI enables analysis of how interaction efficiency—not interaction volume alone—relates to learning outcomes.

Results

This section goes over the analysis from the pre and post proposal activities for the course. In particular, we analysis the outcomes of the assignments (section), and the results of the AI interaction metrics and their correlations (section).

Assignment Outcomes

Table 2: Comparison of Evaluation Scores Between Pre-Proposal and Post-Proposal Assignments

Criteria	**Pre-Proposal Mean (std)**	**Post-Proposal Mean (std)**
Introduction and Motivation	94.0 (9.4)	**96.2** (4.7)
Problem Statement and Objectives	92.4 (10.3)	**93.9** (7.7)
Methodology	**89.9** (11.1)	88.4 (10.4)
Expected Outcomes and Impact	89.5 (12.7)	**91.4** (8.7)
Writing Clarity and Organization	**89.7** (9.9)	88.8 (10.9)
Originality and Creativity	**94.8** (9.4)	93.7 (7.6)
Average Score	**91.8 (8.7)**	**92.2 (7.0)**

Table 2 reports the mean and standard deviation for each rubric criterion. Overall, the average scores show a positive but modest improvement from the pre-proposal to the post-proposal. The most notable gains appear in the Introduction and Motivation,"Problem Statement and Objectives," and "Expected Outcomes and Impact" categories. These increases may reflect students' growing ability to use the AI system to refine the motivation, scope, and anticipated contributions of their projects.

Table 3 summarizes the VisibleAI metrics for authenticity, AI-generated content, pasted content, original content, and AI-revised content. These measures indicate the extent to which proposal text was produced by the student, generated by the AI assistant, or pasted from external sources. Across assignments, students increased both authenticity and original content while reducing AI-generated, pasted, and AI-revised content. This pattern suggests a decreased reliance on the AI tool during the second proposal.

The assignments also included a self-reflection component in which students reported their experience using the Kritik platform. They responded to three questions: *How well did AI help*

Table 3: Summary of AI Insights

Stage	Authenticity	AI Content	Pasted Content	Original Content	AI Revised Content
Pre-Proposal	94%	20%	7%	73%	9%
Post-Proposal	100%	6%	3%	91%	0%

with your assignment? How confident are you in the accuracy of AI-generated content? How much did you learn from using AI? Responses were recorded on a Likert scale ranging from *"Not Effective"* to *"Very Effective."* Most students selected *"Effective"* across all questions, and the ratings remained essentially unchanged between the pre-proposal and post-proposal assignments.

The survey additionally included an open-ended *Written Reflection* prompt. Students' reflections indicate a generally positive experience with the AI-guided assistant, particularly during the early stages of developing their project proposals. Many students reported that the AI helped them brainstorm ideas, clarify concepts, and explore unfamiliar research areas, enabling them to overcome initial barriers and refine their thinking. A recurring theme was the usefulness of the AI's questioning style: by prompting students to rephrase unclear ideas or condense explanations, the assistant encouraged deeper reasoning and improved writing clarity without providing full solutions. Students also valued the AI's support in organizing their proposals, identifying relevant methods or data needs, and improving grammar and structure.

At the same time, several students noted limitations. When they lacked foundational knowledge, the AI's guided approach sometimes felt restrictive, offering questions rather than the concrete information they sought. Others found the assistant helpful for high-level brainstorming but less effective for detailed technical explanations. Overall, students viewed the AI as a valuable tool for ideation, reflection, and refinement, while recognizing that it could not substitute for domain research or expert-level guidance.

AI Interaction Metrics

Figure 1 summarizes the distributions for all AI interaction metrics. Several direct metrics—such as Quality (Q), Structure, and Feasibility—exhibit modest positive trends. The change-based metric Efficiency of Iteration (EOI) also shows an upward trend. The slightly positive mean EOI suggests that, on average, students were able to translate their interactions with the conversational AI into measurable learning gains, indicating that the constrained AI guidance was generally used in a productive manner. The distribution of EOI further reveals distinct patterns of student engagement. A relatively small number of students exhibit low EOI values, corresponding to cases of heavy AI interaction with limited improvement in proposal quality. This suggests inefficient transfer of learning, potentially due to surface-level engagement or misaligned prompting strategies, and highlights the need for additional scaffolding. In contrast, a larger portion of students fall within the higher EOI range, indicating that moderate, goal-directed AI use is associated with substantial improvements in proposal quality. This pattern suggests that, for many students, constrained AI guidance supports efficient learning and meaningful knowledge transfer.

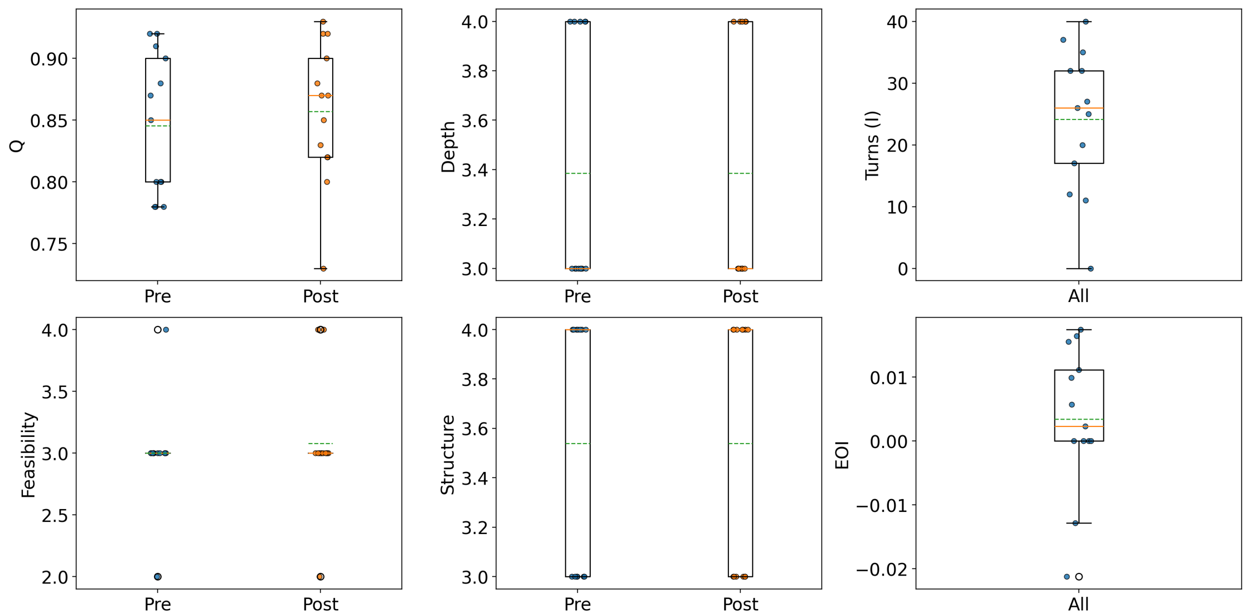

Figure 1: Summary of statistics between all AI interaction metrics. Boxplots for each metric.

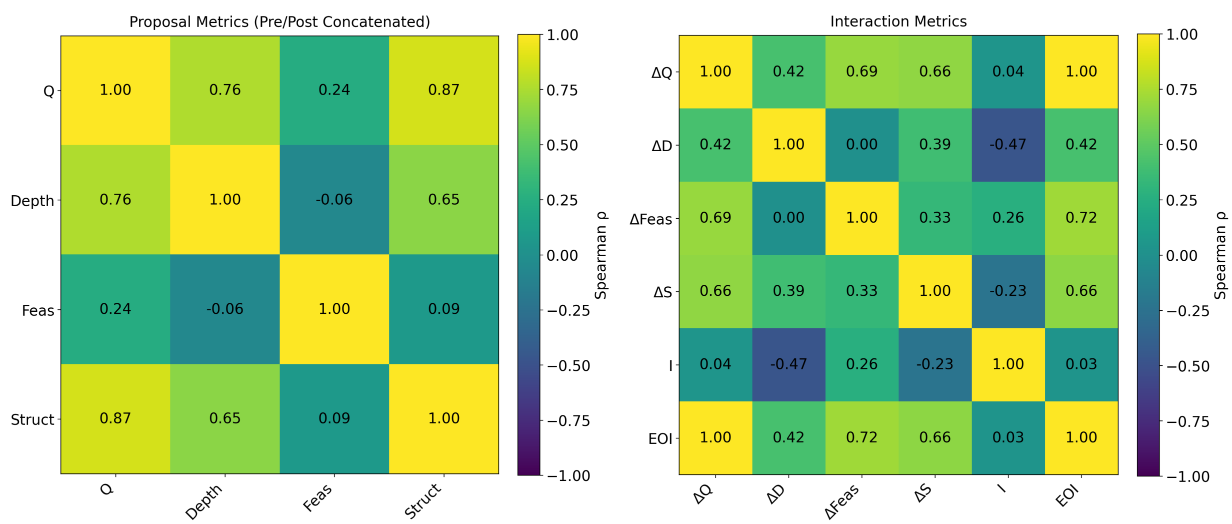

Figure 2: Correlation heatmaps for the AI interaction metrics applicable to a single submission (left), and for the metrics applicable to the change between submissions (right).

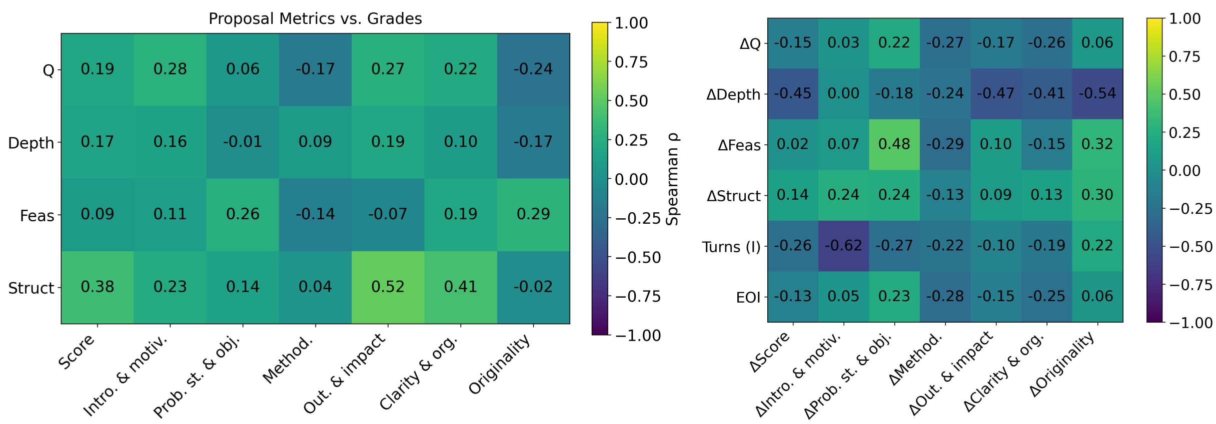

Figure 3: Correlation between direct metrics and assignment scores (left) and change metrics and changes in scores.

Figure 2 (left) presents the correlations among the direct metrics. Quality and Structure are strongly correlated, with a Spearman coefficient of 0.87, followed by Quality and Depth (0.76) and Depth and Structure (0.65). Figure 2 (right) displays correlations among the change metrics. In this case, the strongest relationships appear among ΔQ, Δ Feasibility, and Δ Structure. The high correlation between ΔQ and EOI is expected, given that EOI is derived from ΔQ.

Figure 3 (right) further illustrates the correlations across all change metrics. Examining the correlations with the overall change in score (first column), we observe no strong positive associations. The largest negative correlation occurs between Δ Depth and Δ Score (-0.45). More broadly, the strongest negative relationship is between Student Turns (I) and changes in the Introduction and Motivation criterion. This result is difficult to interpret, however, because the interaction count reflects the total number of turns across both assignments rather than a differential measure. The strongest positive correlation is observed between changes in Feasibility and changes in the Problem Statement and Objectives criterion. This suggests that improvements in project formulation may contribute to increased feasibility of the proposed approach.

Discussion

Some students reported that the AI did not provide the background knowledge they needed and felt “too constrained” when they attempted to brainstorm the broader context of their research topic. Because the goal of the course was to familiarize students with methodologies for solving physics-based problems—rather than to develop expertise across all possible application domains—this feedback suggests that the assignments may have benefited from a two-phase structure: (1) *Ideation* and (2) *Methodology Development*. During the ideation phase, an unrestricted AI system could help students identify, explore, and clarify the concepts necessary to formulate a viable problem statement, including domain ideas outside their prior expertise. During the methodology development phase, a constrained AI guide could then support students in integrating appropriate modeling techniques, reinforcing the concepts they were expected to learn as part of the course.

Based on these insights, an unrestricted AI assistant appears better suited for the early stages of proposal development—particularly topic exploration and selection—whereas a constrained AI guide is more appropriate for planning the methodology and expected outcomes. This distinction points toward a potential restructuring of future assignments.

Several interaction metrics could not be applied in this study because they rely on improvements made during iterative content creation. In future offerings, it may be useful for students to complete the same assignment twice (without changing topics, unlike in the present course), enabling the use of these additional metrics. Moreover, assessing knowledge and quality gains would benefit from the inclusion of control groups (e.g., students working without AI assistance) to more clearly isolate the effects of AI-supported learning.

Although the metrics examined here show some correlations with assignment scores, further investigation with larger cohorts is needed to validate these patterns. An intriguing direction for future work is the design of prompts that correlate directly with rubric-based performance. In other words, can prompts be optimized to produce interaction-derived indices that reliably predict scores across different rubric dimensions?

Conclusion

This study demonstrates that conversational AI, when deliberately constrained to operate as a guide rather than a solution provider, can support meaningful knowledge transfer in a graduate engineering context. Through a two-phase proposal exercise, students showed modest but measurable improvements in proposal quality, particularly in areas related to framing, motivation, and articulating expected outcomes. Interaction-level analyses revealed that students increasingly relied on their own conceptual understanding over time, as evidenced by higher authenticity scores and reduced AI-generated content in the post-proposal stage. While several AI interaction metrics correlated with rubric-based performance, the relationships were nuanced and highlight the need for larger-scale validation.

Overall, distinct patterns in how AI interaction was translated into learning outcomes were observed. Specifically, a subset of students was found to exhibit low EOI, characterized by relatively heavy AI interaction coupled with limited improvement in proposal quality. This pattern suggests that AI guidance was not efficiently converted into learning gains and may be attributed to exploratory overuse, misaligned prompting strategies, or insufficient foundational knowledge rather than a limitation of the AI support itself. In contrast, a larger proportion of students was observed to fall within the high EOI range, where moderate, goal-directed AI use was associated with substantial improvements in proposal quality. Accordingly, this distribution indicates that, for many students, constrained AI guidance was effectively leveraged to support efficient learning and meaningful knowledge transfer rather than dependence on AI outputs.

Student reflections further underscored the value of constrained AI guidance for brainstorming, clarifying ideas, and structuring technical writing, while also revealing limitations when foundational domain knowledge was lacking. These insights suggest that future implementations may benefit from a hybrid approach—using unrestricted AI during early ideation and constrained AI during methodological development—to better align with students' evolving needs.

Overall, the findings point toward the promise of structured, AI-supported assessment frameworks that distinguish conceptual learning from AI interaction expertise. By foregrounding transfer-relevant competencies and emphasizing responsible AI use, this work contributes a practical foundation for designing, evaluating, and scaling AI-guided learning experiences in engineering education.

References

[1] Enkelejda Kasneci, Katharina Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. doi: 10.1016/j.lindif.2023.102274.

[2] Michael Sullivan, Andrew Kelly, and Patricia McLaughlan. Chatgpt in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, 6(1), 2023. doi: 10.37074/jalt.2023.6.1.17.

[3] E. D. Lindsay, S. Otto, and S. Ejsing-Duun. Three frameworks to support engineering education to proactively develop the curriculum with generative ai. In *Proceedings of the Australasian Association for Engineering Education (AAEE) Annual Conference*, 2024.

[4] David H. Jonassen, Johannes Strobel, and Chwee Beng Lee. Everyday problem solving in engineering: Lessons for engineering educators. *Journal of Engineering Education*, 95(2): 139–151, 2006. doi: 10.1002/j.2168-9830.2006.tb00885.x.

[5] Johannes Strobel and Angela van Barneveld. When is pbl more effective? a meta-synthesis of meta-analyses comparing pbl to conventional classrooms. *Interdisciplinary Journal of Problem-Based Learning*, 3(1), 2009. doi: 10.7771/1541-5015.1046.

[6] David N. Perkins and Gavriel Salomon. Transfer of learning. In Torsten Husén and T. Neville Postlethwaite, editors, *International Encyclopedia of Education*, pages 425–441. Pergamon Press, 2 edition, 1992.

[7] Sheri D. Sheppard, Kelly Macatangay, Anne Colby, and William M. Sullivan. *Educating engineers: Designing for the future of the field*. Jossey-Bass, 2008.

[8] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 2023. doi: 10.1007/s40979-023-00146-z.

[9] Behnam Zakeri. Generative ai for project-based assessment in energy and sustainability education. *Research Square*, 2025.

[10] Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012. doi: 10.1111/j.1551-6709.2012.01245.x.

[11] NC State. Kritik. `https://www.kritik.io/`. [Online; accessed 21-Jan-2025].

[12] Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74, 1998. doi: 10.1080/0969595980050102.

[13] Michael J. Prince and Richard M. Felder. Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education*, 95(2): 123–138, 2006. doi: 10.1002/j.2168-9830.2006.tb00884.x.

[14] Christine M. Cunningham and Gregory J. Kelly. Epistemic practices of engineering for education. *Science Education*, 101(3):486–505, 2017. doi: 10.1002/sce.21271.

[15] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC, 2014. ISBN 9780935302356.

[16] Carol A. Chapelle. *Argument-Based Validation in Testing and Assessment*. SAGE Publications, Inc., 2020. ISBN 9781544334486.

[17] Amelia K. Sofjan, Kimberly A. Nguyen, Dhara Surati, and Carine Marette. Assessment of the validity of peer scores and peer feedback in an online peer assessment platform (kritik). *Currents in Pharmacy Teaching and Learning*, 15(4):400–407, 2023. ISSN 1877-1297. doi: https://doi.org/10.1016/j.cptl.2023.04.004. URL `https://www.sciencedirect.com/science/article/pii/S1877129723000709`.

[18] Sidney K. D'Mello and Arthur C. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012. doi: 10.1016/j.learninstruc.2011.10.001.

[19] Vincent Aleven, Bruce M. McLaren, Ido Roll, and Kenneth R. Koedinger. Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(1):205–252, 2016. URL `https://www.cs.cmu.edu/~aleven/Papers/2016/Aleven_etal_IJAIED2016-Helpseeking.pdf`.

[20] Sidney K. D'Mello and Arthur C. Graesser. Confusion. In Reinhard Pekrun and Lisa Linnenbrink-Garcia, editors, *Handbook of Emotions in Education*, pages 289–310. Routledge, 2014.

[21] Yun Bai and Shaofeng Wang. Impact of generative ai interaction and output quality on university students' learning outcomes: a technology-mediated and motivation-driven approach. *Scientific Reports*, 15:24054, 2025. doi: 10.1038/s41598-025-08697-6.

[22] Daniele Agostini, Federica Picasso, and Helga Ballardini. Large language models for the assessment of students' authentic tasks: A replication study in higher education. In *CEUR Workshop Proceedings*, volume 3879, 2024. URL `https://ceur-ws.org/Vol-3879/AIxEDU2024_paper_36.pdf`.