Early Pedagogical Insights from using AI Tutor Agents in a Graduate-level Systems Course

Varad Kulkarni, Nikhil Reddy, and Yogesh Simmhan
Department of Computational and Data Sciences, Indian Institute of Science, Bangalore 560012 India
Email: {varadk, nikhilr, simmhan}@iisc.ac.in

***Abstract*—This paper reports early experiences from teaching an ongoing graduate-level systems course on Cloud Computing using an AI-based Tutor agent as the primary means of instruction in a classroom. We describe the high-level pedagogical workflow we adopt, offer insights of prompts used for tuning the LLMs, propose an analytical framework to quantify student engagement with the AI tutor, and offer preliminary results that reveal consistent patterns of engagement evolution – from broad conceptual exploration to deeper, more focused interactions. This is one of the first such studies on incorporating AI tutors within real classroom environments. These results highlight how structured integration of conversational AI Tutors as a central element of classroom teaching can enhance reflective learning behaviors, provide a reproducible method for studying student engagement, and offer a potentially scalable approach for teaching higher education courses in India, which has a resource gap on high-quality college-level instructors.**

I. Introduction

Conversational AI tools are increasingly being integrated into classrooms, offering new ways for students to learn through inquiry-driven and interactive engagement [1]. Recent studies show that Large Language Models (LLMs) powered agents can enhance classroom participation, promote active inquiry, and support personalized learning experiences by providing contextual explanations and instant feedback [1], [2]. They allow students to engage in self-directed and personalized exploration while receiving adaptive, on-demand guidance from AI tutors.

Despite the growing interest and gradual adoption of AI agents in the classroom, few studies have examined how to design such AI tutoring systems in real classroom and how well they perform in this context, especially in higher education. Most existing works rely on controlled or simulated settings, offering limited understanding of their impact on in-class engagement and inquiry [3]. A methodical study can help drive educational technology and policy, which is particularly critical for a country like India with a population of $1.46B$ at a median age of 29.8[1], poised to reap a demographic dividend but is beset with a lack of sufficient high quality instructors at the college-level. Such pedagogical technologies, if found effective, can help with the educational upliftment of $\approx 260M$ people in the 15–25 higher-education age group.

This study presents one of the first of its kind in-class investigation of designing and using an AI Tutor Agent for an on-going graduate-level *Cloud Computing* course taught at the Indian Institute of Science, Bangalore in the August (Fall) 2025 semester for a cohort of 18 senior undergraduate, Masters and PhD students. An *AI Tutor Agent*, designed by us within Microsoft Teams Copilot, is directly integrated into the *lecture workflow* and configured with weekly module-specific curriculum and guardrails. Students interact with the agent through a chat interface on their laptops – during in-person classes – to explore concepts for the week, clarify doubts, and attempt short, non-graded quizzes. Students submit their chat transcript from these interactions at the end of class, which are automatically evaluated to ascertain their engagement level and feedback returned back to them.

In this short paper, we introduce the key aspects of how we designed and embedded such an AI-driven pedagogical framework within the classroom environment in a structured manner, early experiences on the engagement of students in this setting, and preliminary insights on how their engagement metrics have evolved. A more detailed study is awaiting ethics approval, after which course artifacts (e.g., agents, system prompts, guardrails, surveys, etc.) will be posted online. This goes a step towards an evidence-based template for practically incorporating AI tutor agents within the classroom for graduate and undergraduate education, and particularly for Parallel and Distributed (PDC) courses. Further such studies can have a transformative effect on scaling higher education in India.

II. Background and Related Work

A. Background

LLMs such as OpenAI's GPT and Google's NotebookLM are increasingly being used to support learning and teaching. These models are trained on large corporal of publicly available text, which includes technical and academic content, and can explain concepts, summarize material, and answer questions in natural language. This is helping users interactively obtain information in an intuitive manner. In education, this technology is being used through conversational tools like ChatGPT, which allow students to explore ideas and clarify doubts during study. On the flip side, such LLMs are also causing challenges with large-scale cheating on take home assignments, raising concerns on learning outcomes [4], [5].

LLM services have also introduced dedicated *learning modes* to help users learn more effectively. OpenAI's ChatGPT includes a *Study Mode* that guides learners step by step, asking questions and giving hints instead of directly showing the full answer [6]. This helps students think through problems

[1] https://www.unfpa.org/data/world-population/IN

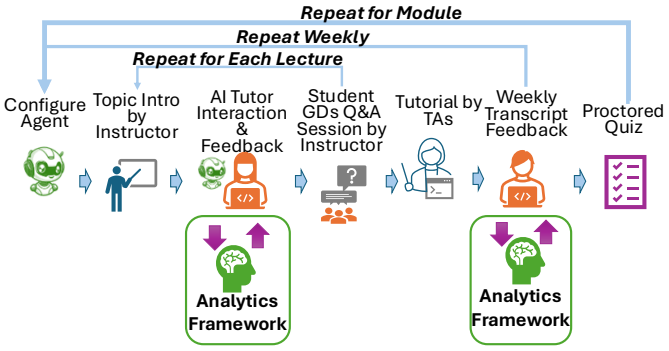


Fig. 1: Agent-driven Pedagogical Workflow

and reflect on their understanding rather than simply copying solutions. Similarly, Google's Gemini offers *Guided Learning* that focuses on explaining concepts gradually and checking user understanding through short, interactive prompts [7]. Such systems are designed for self-paced study and not yet evaluated in live classroom settings. Our work differs by configuring an LLM tutor agent for use within real class sessions to analyze how students interact with it during lectures.

Creating an AI agent typically begins with defining a *system prompt* – a structured set of (English) language instructions that specifies the agent's role, objectives and interaction style. These provide the base context guiding how the LLM interprets the user input and generates responses. Tools such as *Microsoft Copilot Studio* allow developers to build multi-turn conversational agents by embedding documents, URLs or datasets directly into the prompt context and controlling parameters such as tone and reasoning depth [8]. Configuring such agents is an iterative process involving prompt refinement, and evaluation to achieve consistency and factual reliability. We take this approach in this study.

B. Related Research

The integration of LLMs into educational practice has gained great interest in recent years. Systematic reviews have examined their pedagogical potential to enhance inquiry-driven and interactive learning [9], [10]. These works consistently highlight LLMs' capacity to provide context-sensitive explanations, adaptive feedback and personalized learning support, positioning them as extensions of human tutoring rather than simple information retrieval systems. Chu et al. [11] present a detailed survey of LLM-based educational agents and their role in fostering metacognitive engagement. More recent theoretical frameworks, such as the ICAP-based mapping of cognitive engagement proposed by Shah [12], emphasize the need for structured evaluation of dialogue-based learning. However, most prior research focuses on conceptual or experimental prototypes, with limited exploration of how these systems behave in real classroom environments, which is a novel focus of our study.

A growing collection of works is investigating the use of LLMs to simulate or augment classroom instruction. Zhang et al. [3] introduce the *SimClass framework*, where multiple AI agents emulate teachers and students to replicate classroom discourse patterns. Li et al. [13] extended this idea through human–LLM collaborative teaching environments that balance automation with guidance. These studies demonstrate the feasibility of AI-mediated classroom orchestration but primarily operate within simulated or hybrid experimental setups. LLMs have also been explored as feedback-oriented tools, e.g., for grading design assignments that lead to improved feedback clarity but with limitations in contextual reasoning [14]. We apply several of these techniques in-the-wild, in a real classroom setting, using LLMs both as a primary instructor and for feedback, and report our initial experiences.

III. AI Tutor Framework

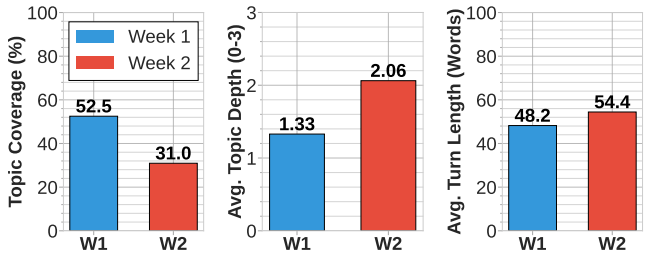
A. Pedagogical Workflow

Our pedagogical framework incorporating the AI Tutor is shown in Fig. 1. It integrates LLM-based tutor agents into a structured, repeatable instructional workflow spanning the full duration of the course. The system is implemented within the Microsoft Teams Copilot environment and linked to Moodle, a widely used Learning Management Systems (LMS). We also use Function-as-a-Service (FaaS) workflows running on the cloud to process transcripts, compute engagement metrics, and securely store analytical reports.

The bootstrap for the workflow is the instructor designing the *course curriculum* for the entire semester, split into different *modules*, with each module spanning multiple *topics*, one per week, and the concepts and learning objectives defined for each topic. Each week has two *lecture* sessions and one hands-on *tutorial* session. The workflow operates at three repeating granularities. A base agent is refined at the *weekly topic level*, where the agent prompt is configured with the set of topics for the lectures and tutorial this week. The same AI Tutor instance is used for the interactions for this week's topics.

At the *lecture level*, each in-person session follows a cycle of instructor-led *introduction* to the topic, then *student interactions with the AI Tutor* in-class for a self-paced approach to soliciting details for topics, followed by an in-class *peer discussion* to help firm up the concepts through discussion, and lastly, an instructor led question and answer session. At the end of the chat interactions, each student submits their chat transcript using a course form for personalized feedback. At the *module level*, comprising of 2–4 weeks of activities, we have a proctored in-class unaided quiz to evaluate their conceptual understanding.

The base AI Tutor agent is instantiated with three structured components: (1) *System-level prompts* that define instructional goals, response style, and reasoning scope, along with guardrails to not to deviate from the goals of instructional support and strive to be accurate rather than agreeable; (2) *Pedagogical alignment prompts* for inquiry-based and scaffolded learning based on Knowledge–Learning–Instruction (KLI) framework [15]; and (3) *Topic knowledge base* for the week's activities, derived from the overall curriculum, with learning outcomes based on Bloom's taxonomy. The first two components are common to all weekly agents for the course.



(a) Topic Coverage (b) Avg. Topic Depth (c) Avg. Turn Length Per Topic

Fig. 2: Median of the students engagement metrics for each of the two weeks of activities being compared.

B. Engagement Analytics

Our Analytics framework (Fig. 1) quantitatively evaluates student's engagement with the AI Tutor based on the classroom chat transcripts uploaded by them after each lecture. Each transcript captures a multi-turn conversation sequence for various topics, and agent-invoked actions such as quizzes or summaries. The framework transforms these raw dialogue records into structured indicators of engagement and learning behavior, that is shared with the student to help them improve their interaction and learning outcomes.

Each student's raw dialogue file is passed to our evaluation engine. The engine itself is an AI agent configured through a structured system prompt that specifies how to analyze the transcript in a consistent and unbiased manner across modules [16]. The evaluation agent follows a schema-driven analytical process that enforces fixed interpretation rules and standardized output fields. For every transcript, it produces a structured report that summarizes both behavioral and conceptual dimensions of engagement. The resulting output is designed to be interpretable by students and instructors alike, while remaining compatible with downstream aggregation and visualization for higher-level analysis.

Three primary metrics are used to capture engagement dynamics and evaluated by our agent, at the granularity of each subtopics specified for the week's topic [17].

- *Topic Coverage:* Measures the proportion of canonical subtopics from the module that the student actively engaged with during the session, reflecting the breadth of conceptual exploration.
- *Topic Depth:* Represents the depth score of each subtopic discussed in the module, indicating how thoroughly students explored a concept. Each subtopic is rated on a four-level ordinal scale: 0 – Briefly mentioned, 1 – Basic question asked, 2 – Explored with follow-ups or comparisons, and 3 – Examined in depth through reasoning or clarification.
- *Turn Length:* Calculates the number of words per student message within a subtopic, serving as an indicator of elaboration and reflective effort. Longer turns typically suggest greater reasoning and conceptual articulation.

A subset of these raw transcripts were manually evaluated to ensure consistency with the three individual AI-evaluated metrics. Besides providing feedback to individual students, each transcript's metrics are also aggregated across students and topics and share with the instructor to give a sense of the engagement during each lecture. This helps the instructor intervene during the initial primer or end of lecture question session. It also enables comparative analysis of the engagement evolution over time. This unified schema provides a consistent foundation to study shifts in inquiry behavior and pedagogical effectiveness in AI-mediated learning environments.

IV. Preliminary Results

The proposed AI Tutor and evaluation framework is deployed and being evaluated in a graduate-level Cloud Computing course at IISc, with both senior undergraduates from computer science and graduate students from multi-disciplinary backgrounds enrolled. This study is limited to the first module of instruction, and offers an early preview of experiences.

Activities from two weeks of the first module are reported and compared: *Virtualization and Container Runtimes* (20 subtopics) and *Cloud Service and Deployment Models* (21 subtopics) [2]. Each week's agent was instantiated following the design process described in Section III-A and the transcripts from student interactions with the AI Tutor was analyzed using the evaluation procedure in Section III-B. Participation was restricted to the 18 students formally enrolled in the course (and who continue to be enrolled as of writing, at the end of 8 weeks), and user consent was obtained for this study. All interaction logs were anonymized before analysis.

Engagement was assessed across three aggregated metrics: topic coverage, average topic depth (mean topic depth across all topics for the module), and average turn length per topic. These metrics collectively represent the depth of reasoning, coverage of exploration, and behavioral engagement. Their median values for each week's transcripts are visualized in Figure 2.

1) Topic Coverage and Depth: Figures 2a and 2b summarize the evolution of topic coverage and depth across the two instructional weeks. A contraction in topic coverage is observed, decreasing from approximately 52.5% of canonical topics in Week 1 to 31.0% in Week 2. Concurrently, the average topic depth increased from 1.33 to 2.06 on a three-point ordinal scale. This indicates a shift from broad conceptual exploration to more concentrated inquiry on selected topics, as students clarified core uncertainties. Lower topic coverage in Week 1 could be attributed a combination of time constrain during in-class environment; participation from home (dorm) in a flipped-class format was limited, indicating motivational barriers to off-class learning.

2) Turn Length: In Figure 2c, the average turn length per topic increased modestly from 48.2 to 54.4 words between the two weeks. This indicates that student responses became slightly longer and more elaborate over time. The increase likely reflects growing familiarity with the conversational

[2]Since these were the initial part of the course, each "week" of activity spilled over into 1.5 weeks, i.e., 3 lectures each. Results are reported over this 3-week period.

format and aligns with the rise in topic depth, indicating a shift towards more reflective and information-rich dialogue.

3) Inter-Metric Correlations: The three engagement metrics together reveal a consistent trend. As topic coverage narrowed by roughly 41%, both topic depth ($+55\%$) and average turn length ($+13\%$) increased. This inverse relation between coverage and depth indicates that students gradually shifted from broad exploration toward more focused, conceptually dense exchanges. The positive alignment between depth and average turn length further suggests that deeper reasoning was accompanied by longer, more reflective responses. Overall, these patterns point to a transition from exploratory inquiry to deliberate, focused engagement as the course progressed.

Taken together, these early findings show students adapting their use of the AI Tutor as the course progressed. Early sessions were characterized by wide exploration across topics, while later ones show more targeted and detailed engagement.

V. Conclusion

This study presents an integrated approach for analyzing student engagement with LLM-based tutor agents in a real classroom setting, acting as the primary instructor, guided and complemented by human instructors and TAs. It offers early insights into an AI-driven pedagogical framework for configuring and integrating instructional agents aligned with inquiry-based learning, along with an analytical framework to quantify engagement through dialogue-derived metrics.

The classroom deployment of LLM-based tutor agents revealed both pedagogical benefits and practical challenges. The framework successfully captured engagement trajectories, showing a shift from broad exploration to deeper inquiry, yet engagement alone does not guarantee learning effectiveness. A natural next step is to correlate these behavioral indicators with outcome-based measures such as graded assessments and in-class evaluations to assess how AI-mediated interactions translate into actual learning gains. These are part of our ongoing work as the semester progresses. The study also showed that AI tutoring works particularly well for information-rich subjects like Cloud Computing, where abundant public content supports factual grounding, limiting the need for instructor supplied context. Early student feedback also indicate a high level of satisfaction with the course, with only rare instances of hallucinations reported, though a more detailed analysis of student surveys is awaiting ethics approval.

Future work will also expand this framework by integrating *Agentic AI* workflows capable of autonomous reasoning, tool use, and adaptive feedback to enhance evaluation accuracy and contextual awareness, and also lead tutorial sessions for systems courses that are particularly complex due to their hands-on nature. We plan to correlate engagement metrics with assessment results and collect student feedback on learning effectiveness. Some students reported hallucinations in the tutor agents' responses, we will include a report on factual errors and hallucinations. These extensions aim to evolve the framework into a scalable, outcome-driven model for personalized and accountable AI-assisted education.

Acknowledgements

We thank Profs. Viraj Kumar and Deepak Subramani from IISc for their pedagogical inputs and feedback.

References

[1] R. Abu Khurma, O. Al-Kurdi, M. Alshurideh *et al.*, "Ai chatgpt and student engagement: Unraveling dimensions through prisma analysis for enhanced learning experiences," *Education and Information Technologies*, 2024.

[2] T. K. F. Chiu, X. Zhou *et al.*, "Artificial intelligence in education: A systematic review of opportunities, challenges, and implications," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100204, 2023.

[3] Z. Zhang, D. Zhang-Li, J. Yu, L. Gong, J. Zhou, Z. Hao, J. Jiang, J. Cao, H. Liu, Z. Liu, L. Hou, and J. Li, "Simulating classroom education with llm-empowered agents," arXiv, Preprint arXiv:2406.19226, 2024. [Online]. Available: https://arxiv.org/abs/2406.19226

[4] S. Leaton Gray, D. Edsall, and D. Parapadakis, "Ai-based digital cheating at university, and the case for new ethical pedagogies," *Journal of Academic Ethics*, pp. 1–18, 2025.

[5] "Revealed: Thousands of uk university students caught cheating using ai," The Guardian, June 2025, https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey.

[6] OpenAI, "Introducing study mode in chatgpt," https://openai.com/index/chatgpt-study-mode/, 2025, accessed: 2025-08-10.

[7] Google, "Guided learning in gemini: From answers to understanding," https://blog.google/outreach-initiatives/education/guided-learning/, 2025, accessed: 2025-08-10.

[8] Microsoft Corporation, "Microsoft copilot studio documentation: Build and configure copilots," https://www.microsoft.com/en/microsoft-copilot/microsoft-copilot-studio, 2024, accessed: 2025-10-14.

[9] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?" *Intl. Journal of Educational Technology in Higher Education*, vol. 16, no. 1, 2019.

[10] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. L. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, 2023.

[11] B. Dong, J. Bai, T. Xu, and Y. Zhou, "Large language models in education: A systematic review," in *International Conference on Computer Science and Technologies in Education (CSTE)*, 2024.

[12] R. Techawitthayachinda and R. Iriya, "Automatic assessment of active learning in online discussions with large language models," in *International Conference on Artificial Intelligence in Education Technology*. Springer, 2024, pp. 34–42.

[13] X. Li, R. Chen, and J. Gao, "Human–ai collaborative teaching: Designing classroom experiences with large language models," *Computers & Education*, 2024.

[14] Q. Huang, T. Willems, and K. W. Poon, "The application of gpt-4 in grading design university students' assignments and providing feedback: An exploratory study," arXiv, Preprint arXiv:2409.17698, 2024. [Online]. Available: https://arxiv.org/abs/2409.17698

[15] K. Koedinger, A. T. Corbett, and C. A. Perfetti, "The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive science*, vol. 36 5, 2012.

[16] V. Kulkarni, N. Reddy, and Y. Simmhan, "Pedagogy meets ai & systems: Towards orchestrating tutor agents in moodle using faas," in *HiPC Student Research Symposium*, 2025, to appear.

[17] N. A. Flanders, *Analyzing Teacher Behavior*. Reading, MA: Addison-Wesley, 1970.