Learning in Blocks: A Skill-Driven Framework for Assessment-Centered Agentic AI in Education

Nicy Scaria* NICYSCARIA@IISC.AC.IN
Silvester John Joseph Kennedy* SILVESTERJKENNEDY@GMAIL.COM
Deepak Subramani DEEPAKNS@IISC.AC.IN
Computational and Data Sciences
Indian Institute of Science, Bengaluru

Abstract

Generative AI is increasingly being positioned as a tutoring and learning-support technology, often through chatbot-based interfaces. While flexible, this design can make it difficult to constrain system behavior, align interactions with curriculum goals, and connect learner activity to demonstrated skill performance. This position paper proposes *Learning in Blocks*, a skill-driven framework for AI-supported learning in which learners engage with structured assessments and practice tasks rather than open-ended conversation as the primary interface. The framework organizes learning into blocks of target and prerequisite skills. Bounded pedagogical agents support assessment generation, assessment evaluation, diagnostic recommendation, spaced review, and mastery-based progression. We argue that responsible AI in education should center structured evidence of learner skill. By integrating adaptive learning, formative assessment, multifaceted evaluation, spaced repetition, and mastery learning into a single loop, Learning in Blocks offers a design pattern for more transparent, auditable, and pedagogically aligned AI-supported learning.

Keywords: Adaptive Learning; Agentic AI; Mastery-Based Progression; Responsible AI in Education; Skill-Driven Learning

1. Introduction

Generative AI is increasingly positioned as tutoring and learning-support technology, with major systems emphasizing personalized help, guided learning, and conversational support (Team et al., 2024). While promising, this direction raises concerns around reliability, pedagogical alignment, learner safety, and responsible classroom use (Kasneci et al., 2023; Zheng et al., 2023; Li et al., 2025). These concerns are heightened when learning depends on open-ended conversation, where system behavior can be difficult to constrain, inspect, and connect to evidence of learning.

This paper argues that AI-supported learning should not be organized only around chatbot-based personal tutors. We propose *Learning in Blocks*, a skill-driven framework in which learners engage primarily with structured assessments and practice tasks rather than free-form dialogue. Bounded pedagogical agents generate or select assessments, evaluate learner artifacts, diagnose skill gaps, recommend targeted practice, schedule spaced review, and support mastery-based progression. The framework draws on adaptive learning, formative assessment, rubric-based evaluation, spaced repetition, and mastery learning. It integrates these principles into a single learning loop where personalization is driven by skill evidence and instructional decisions are grounded in curriculum-defined rubrics, review schedules, and progression criteria.

* These authors contributed equally.

2. Toward Skill-Driven AI Learning Systems

Adaptive learning systems provide important foundations for modeling learner state and sequencing instruction. Knowledge tracing, item response theory, and intelligent tutoring systems estimate learner proficiency and select next activities, but many rely on item-level correctness, selected-response tasks, or narrow interaction traces (Corbett and Anderson, 1994; VanLehn, 2011; Piech et al., 2015; Abdelrahman et al., 2023; Shen et al., 2024; Imamah et al., 2024). This creates a challenge for skill-driven domains such as STEM, where mastery often requires applied performance across explanations, derivations, diagrams, code, or transfer tasks rather than correctness on isolated items.

Recent work on rubric-based and LLM-enabled evaluation suggests that AI systems can increasingly evaluate open-ended artifacts using multidimensional criteria (Hashemi et al., 2024; Zheng et al., 2023; Li et al., 2025). Multi-agent and debate-style approaches further suggest that reliability may be improved by comparing or synthesizing multiple model perspectives (Du et al., 2024). These developments make multifaceted evaluation more practical than in earlier adaptive systems. However, evaluation alone is insufficient unless rubric judgments are connected to instructional decisions.

This connection is central to formative assessment and feedback. Assessment improves learning when evidence of performance is used to adapt subsequent instruction, and effective feedback helps learners close the gap between current and desired performance (Black and Wiliam, 1998; Nicol and Macfarlane-Dick, 2006; Shute, 2008). Deployed systems such as ASSISTments show that connecting assessment evidence to instructional support is feasible at scale (Heffernan and Heffernan, 2014), but their evaluation remains grounded in item-level correctness and scaffolding metrics, and many recommendation mechanisms remain tied to item tags or topic sequences rather than multifaceted skill evidence.

The framework also draws on spaced repetition and mastery-based progression. Distributed practice supports long-term retention (Cepeda et al., 2006); in this framework, spaced repetition is realized through varied skill-aligned review rather than verbatim item repetition. Mastery learning argues that learners should progress after demonstrating sufficient competence rather than after fixed time or content completion (Bloom, 1968; Guskey, 2010). Recent work further highlights mastery thresholds as an important design and evaluation question (Matayoshi et al., 2025). For skill-driven learning, prior skills should be revisited through diverse assessment formats, and progression should depend on evidence that prerequisite skills are stable enough to support future learning.

Together, these strands provide components for an integrated AI learning system, but they are often treated separately. Traditional adaptive systems model learner state but often rely on narrow evidence; LLM-based evaluation can assess richer artifacts but does not define how learners should progress; feedback, spaced practice, and mastery learning provide instructional principles but are not always unified around skill diagnosis. At the same time, chatbot-centered educational AI raises concerns around reliability, appropriateness, and learner safety when students interact directly with generative models in unrestricted ways (Kasneci et al., 2023). The integration proposed here closes the loop from multifaceted evaluation of learner artifacts through diagnostic recommendation to spaced review and mastery-based progression within a single skill-driven cycle, while keeping AI actions structured, auditable, and curriculum-aligned.

3. The Learning in Blocks Framework

Learning in Blocks organizes learning around skill-centered blocks, structured assessments, pedagogical agents, and mastery-based progression. The framework is designed to make progression depend on demonstrated mastery of target skills.

Each learning block, CB_t, is organized around prerequisite and target skills. Let $S_t = \{s_{t,1}, s_{t,2}, \ldots, s_{t,n}\}$ denote the skills targeted in CB_t, and let P_t denote the prerequisite skills required for those targets. A skill may involve conceptual, procedural, representational, explanatory, computational, or transfer-oriented performance. For each skill $s_{t,i}$, the system assigns one or more skill-aligned assessments $A_{t,i}$ whose format depends on the evidence needed to evaluate that skill, such as worked solutions, explanations, diagrams, code traces, or transfer problems. Progression from CB_t to CB_{t+1} requires sufficient mastery evidence over S_t. Once mastered, skills in S_t are added to the learner's prerequisite skills and may be included in P_{t+1} or carried forward for review in later blocks. Evaluation is conditioned on cumulative exposure, so learners are not penalized for demonstrating skills beyond S_t; such evidence can inform enrichment or personalization of later blocks. For example, a quadratic equations block could assess method selection, algebraic execution, root interpretation, and graphical connection through a worked solution, error-diagnosis task, and graph-interpretation item. If a learner correctly solves a quadratic equation but misinterprets the meaning of its roots, the evaluation profile would show strength in algebraic execution but weakness in root interpretation. The recommendation step would then prioritize practice on interpreting roots in context, while algebraic execution could be carried forward for later spaced review rather than remediated immediately.

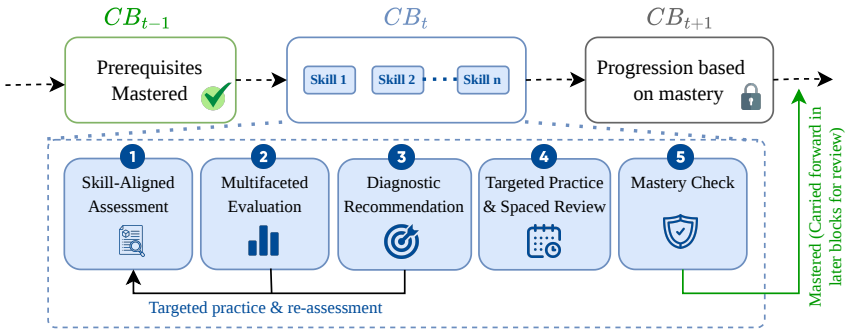

Figure 1: Learners progress through blocks via a recurring loop of assessment, evaluation, recommendation, targeted practice, and mastery checks.

The system can be implemented as a set of bounded pedagogical agents, each responsible for a constrained part of the learning loop. In one implementation, an assessment agent generates or selects tasks aligned to target skills; an evaluation agent applies predefined rubric dimensions to learner artifacts; a diagnosis agent maps rubric evidence to skill strengths and gaps; a recommendation agent selects targeted practice from approved activities; a review agent schedules previously mastered skills for spaced review based on observed skill stability; and a progression agent determines whether mastery evidence is sufficient to move to the next block. Boundedness is operationalized through explicit inputs,

permitted actions, and output schemas. For example, assessment generation is constrained by the skill map, evaluation by predefined rubric dimensions, recommendation by approved practice activities and skills, review by the learner's prior mastery evidence, and progression by mastery criteria. These constraints make system decisions inspectable and auditable.

As shown in Figure 1, these stages form a recurring loop. The learner completes a skill-aligned mastery check, receives rubric-based evaluation and diagnostic recommendations, engages in targeted practice and spaced review, and progresses only when mastery evidence is sufficient. Mastered skills are carried forward into later blocks for periodic review to support retention. The review set is updated as new assessment evidence is observed, so spaced review reflects both prior mastery and later changes in skill stability.

This design supports teacher oversight and safer classroom deployment. Because system actions are grounded in predefined skills, assessments, rubrics, recommendation rules, and mastery criteria, teachers or platform designers can inspect, override, or calibrate key decisions. In this way, *Learning in Blocks* connects structured AI-driven instruction with continuous evaluation, pedagogical alignment, and classroom readiness, while reducing the risks associated with direct open-ended learner interaction with generative models.

4. Design Principles and Limitations

Learning in Blocks is guided by four design principles. First, the framework adapts to demonstrated skill performance rather than treating content completion as evidence of learning. Second, learner interaction is assessment-centered, with structured tasks producing evidence that can be evaluated and acted on. Third, system actions remain curriculum-aligned through predefined skill maps, rubrics, recommendation rules, review schedules, and mastery criteria. Fourth, the framework supports teacher oversight through inspectable and adjustable assessment, recommendation, review, and progression decisions. Together, these principles allow the system to personalize learning while keeping instructional decisions grounded in structured evidence and curriculum-defined constraints.

The framework also has limitations that should shape future work. Most importantly, the current paper presents Learning in Blocks as a design framework without preliminary deployment results, learner outcome data, or teacher usability evidence. As a result, it should be read as a framework for structuring responsible AI-supported learning rather than as evidence that assessment-centered agentic AI improves learning outcomes over existing tutoring or adaptive-learning systems. Designing skill maps and rubrics requires substantial domain expertise, and LLM-based evaluation must be validated for reliability, bias, and consistency across learner groups. Mastery thresholds also require calibration. If thresholds are too low, learners may progress with unstable prerequisite skills; if they are too high, learners may become stuck, bored, or disengaged despite being ready for productive struggle in the next block. Responsible deployment therefore requires threshold tuning, teacher override, and mechanisms for enrichment or alternative pathways when learners repeatedly fail to progress. Finally, because recommendations influence what learners practice and their progress, they should be audited for fairness across learner groups. Such audits should examine whether comparable learner evidence leads to comparable evaluations, recommendations, review schedules, and progression decisions. They should also check whether any

group is systematically assigned less challenging material, fewer advancement opportunities, longer remediation paths, or lower-quality practice.

5. Conclusion

Learning in Blocks offers a skill-driven design pattern for AI-supported learning centered on assessment, recommendation, review, and mastery-based progression. Learners engage with structured tasks, while bounded pedagogical agents evaluate artifacts, diagnose skill gaps, recommend practice, schedule review, and guide progression. By grounding these actions in skill maps, rubrics, review schedules, and progression criteria, the framework aims to make personalization transparent, auditable, and pedagogically aligned. Responsible AI in education should support evidence-based decisions about learner skill, targeted practice, review, and readiness to progress.

References

Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

Benjamin S Bloom. Learning for mastery. instruction and curriculum. regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1. *Evaluation comment*, 1(2):n2, 1968.

Nicholas J Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354, 2006.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Thomas R Guskey. Lessons of mastery learning. *Educational Leadership*, 68(2):52–57, 2010.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, 2024.

Neil T Heffernan and Cristina Lindquist Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International journal of artificial intelligence in education*, 24(4):470–497, 2014.

Imamah, Umi Laili Yuhana, Arif Djunaidy, and Mauridhi Hery Purnomo. Enhancing students performance through dynamic personalized learning path using ant colony and item response theory (ACOIRT). *Computers and Education: Artificial Intelligence*, 7:100280, 2024.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025.

Jeffrey Matayoshi, Eric Cosyn, Hasan Uzun, and Eyad Kurd-Misto. Using a randomized experiment to compare mastery learning thresholds. *Journal of Educational Data Mining*, 17(1):308–336, 2025.

David J Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879, 2024.

Valerie J Shute. Focus on formative feedback. *Review of educational research*, 78(1):153–189, 2008.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36:46595–46623, 2023.