
BUILDING AI COMPANIONS THAT PRIORITISE LEARNING OVER PERFORMANCE

Hassan Khosravi*
The University of Queensland
St Lucia, QLD 4072, Australia
h.khosravi@uq.edu.au

Dragan Gašević
Monash University
Clayton, VIC 3800, Australia
dragan.gasevic@monash.edu

Shazia Sadiq
The University of Queensland
St Lucia, QLD 4072, Australia
shazia@eecs.uq.edu.au

Lixiang Yan
Tsinghua University
Haidian District, Beijing 100084, China
lixiangyan@tsinghua.edu.cn

Jason M. Lodge
The University of Queensland
St Lucia, QLD 4072, Australia
jason.lodge@uq.edu.au

Jason M. Tangen
The University of Queensland
St Lucia, QLD 4072, Australia
j.tangen@uq.edu.au

Paul Denny
The University of Auckland
Auckland 1010, New Zealand
p.denny@auckland.ac.nz

Kristen DiCerbo
Khan Academy
Mountain View, CA 94041, USA
kristen@khanacademy.org

Simon Buckingham Shum
University of Technology Sydney
Ultimo, NSW 2007, Australia
simon.buckinghamshum@uts.edu.au

Ryan S. Baker
Adelaide University
Adelaide, SA 5005, Australia
ryan.baker@adelaide.edu.au

ABSTRACT

Large language models (LLMs) are rapidly transforming knowledge work by improving the quality and efficiency of tasks such as writing, coding, and data analysis. However, their growing use in education has exposed a learning–performance paradox: while they can enhance short-term task performance, they may also undermine genuine learning, including cognitive growth, knowledge transfer, and metacognitive development. This paper addresses the question of how artificial intelligence should be designed and used to support learning rather than merely improve immediate outputs. We introduce the concept of AI learning companions, defined as adaptive, pedagogically informed, LLM-powered agents designed for integration into learning environments. We propose a framework for their design built on three interrelated foundations: a pedagogical foundation focused on how students learn with AI, an adaptive foundation focused on how AI learns about students, and a responsible design foundation ensuring systems remain transparent, accountable, inclusive, and secure. The framework is illustrated through five case studies spanning diverse educational contexts, levels, and tool designs, revealing both the promise and current limitations of existing tools. We conclude that there is a necessary shift away from LLMs designed for task-oriented performance, and beyond simply prompting them to act as tutors, toward deliberately developed AI learning companions that are pedagogically sound, adapt to their learners, and foster durable understanding, metacognitive growth, and learner agency.

Keywords Generative AI · AI in Education · Adaptive Learning · Learning companions

1 Introduction

Generative artificial intelligence (AI), particularly large language models (LLMs), is rapidly transforming the nature of knowledge work by enhancing productivity and reshaping professional practices. Unlike earlier forms of automation that targeted repetitive, low-cognition tasks, LLMs extend automation into domains requiring advanced cognitive skills such as writing, analysis, programming, and decision-making [23]. This marks a profound shift from automating routine labour to augmenting complex, knowledge-intensive work.

This shift aligns with the paradigm of co-intelligence or *hybrid intelligence*, in which human and artificial intelligences are deliberately combined to exceed the capabilities of either alone [49]. AI systems contribute speed, scale, and precision to data-intensive processes, while humans provide contextual awareness, ethical reasoning, and creative synthesis [31]. A growing body of field studies substantiates these benefits, demonstrating substantial time savings and performance boosts across sectors. For instance, studies on knowledge workers using Microsoft 365 Copilot report significant reductions in time spent on email and document drafting [35], while engineers using GitHub Copilot completed coding tasks 56% faster [83]. This productivity leap is reinforced by experimental evidence in professional writing tasks, where generative AI was found to reduce task completion time by 40% while simultaneously increasing output quality by 18% [74]. Similarly, AI augmentation has been shown to increase customer service agent efficiency by 14% and help less experienced workers narrow the skills gap with experts by enabling newer employees to move up the productivity curve more quickly than traditional training allows [18, 30].

LLMs are rapidly becoming embedded in students' everyday study practices, supporting tasks such as summarising, reasoning, writing, coding, and analysis [25]. As education systems shift from prohibition towards more structured adoption, many students remain uncertain about how to use AI effectively for learning, with some voicing concerns about cheating themselves of genuine understanding [25]. These concerns are consistent with UNESCO and OECD warnings about the widening gap between the pace of technological change and the pedagogical frameworks needed to guide its effective use in education [104, 77]. They also echo longstanding findings from the learning sciences that improvements in immediate task performance do not necessarily lead to durable learning [99, 118]. While recent meta-analyses report that students using ChatGPT or similar systems show improved performance on assignments and exams [32, 44], these findings must be interpreted with caution: the studies on which they draw largely measure short-term, scaffolded performance rather than durable learning, and have been critiqued as an effect in search of a cause [110].

The distinction between performance and learning is not merely semantic. It reflects what the literature terms the learning-performance paradox [99, 118]: the well-documented phenomenon whereby AI tools can enhance short-term task outputs while simultaneously undermining durable learning, including cognitive growth, knowledge transfer, and metacognitive development. This paradox arises from a well-documented cognitive mechanism: cognitive offloading, defined as the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand [91]. When students rely on AI for sensemaking, planning, monitoring, and evaluation, they reduce engagement in the very processes that make self-regulated learning effective [39, 119]. This matters because durable understanding emerges through grappling with complexity rather than bypassing it [51]. The most compelling empirical evidence for this comes from a large randomised experiment in high school mathematics, in which students given access to an AI assistant showed improved problem-solving during learning but suffered significant harm to their durable, independent learning once the AI was removed [10]. The same pattern has been observed in misconception correction: Corbett et al., [27] found that personalised AI dialogue produced significantly larger immediate belief reductions than traditional textbook refutation, but by the two-month follow-up the two conditions had converged, suggesting that the engagement and confidence generated by AI interaction did not translate into durable learning advantages. AI-supported tasks can thus produce stronger artefacts without equivalent learning gains, alongside reduced self-regulatory monitoring [29], increased dependence on AI assistance, and a form of metacognitive laziness in which learners abdicate the cognitive and metacognitive effort required for deep understanding [39]. Optimising for immediate task completion therefore comes at the cost of cultivating the processes that support durable learning [118, 99].

These findings illuminate a fundamental tension. While LLMs show clear potential for supporting hybrid intelligence and enhancing productivity in professional contexts, education requires a different orientation, one that values process over product, prioritises cognitive growth over immediate results [13, 99], and fosters metacognitive development through reflection, self-regulation, and strategic awareness of one's own learning processes [120]. This tension raises a central question: how should AI be designed and used to cultivate learners' cognitive and metacognitive abilities rather than merely support short-term performance?

In response, prompt-level approaches have emerged to steer LLMs towards more pedagogically productive interaction. Commercial systems now include study-oriented modes that use questioning, step-by-step guidance, and checks for understanding rather than simply returning answers. Although empirical studies indicate that prompt-based guardrails

can mitigate many of the negative effects associated with direct, unstructured LLM use [10], their effects on learning are often near null, and they are unlikely to match the learning benefits historically associated with Intelligent Tutoring Systems (ITSs) [15]. This is because prompt-based approaches remain limited in three important respects: first, they are typically reactive rather than proactive; second, they are largely stateless in relation to learners' prior knowledge, goals, and misconceptions; and third, they operate at the level of specific tasks rather than supporting broader learning trajectories across the curriculum. By contrast, ITSs and adaptive learning systems have long addressed these challenges through student modelling and knowledge tracing [1], and through instructional adaptation engines that personalise feedback, scaffolding, and learning pathways [4] with strong evidence of learning benefits [117]. Yet these systems have been difficult to scale because they rely heavily on manually authored content, feedback rules, and pedagogical pathways [34]. A significant gap therefore remains in how to integrate the conversational flexibility and scalability of LLMs with the learner modelling and pedagogical adaptivity of ITSs to create AI-powered educational tools that genuinely support learning.

Building on decades of research in the learning sciences, adaptive and intelligent tutoring systems, and human-centred approaches to AI, we introduce the concept of AI learning companions: adaptive, pedagogically informed, LLM-powered agents designed for integration into diverse educational environments and explicitly engineered to prioritise durable learning over short-term performance. We propose a framework for their design built on three interrelated foundations. The pedagogical foundation addresses how students learn with AI, fostering cognitive engagement, metacognitive awareness, and a sense of agency throughout the learning process. The adaptive foundation addresses how AI learns about students, modelling cognitive, affective, and behavioural data to tailor feedback, adapt instruction, and sustain personalised, data-informed feedback loops. The responsible design foundation ensures that AI companions remain transparent, inclusive, accountable, and secure, upholding human oversight, equity, and trust. In what follows, we first conceptualise learning companions by synthesising insights from the relevant literature to clarify their defining characteristics and affordances. We then elaborate on the design and implementation of such companions across the three foundations, detailing how each contributes to building AI-based learning companions, and illustrate the framework through five case studies spanning diverse educational contexts, levels, and tool designs.

2 Conceptualising Learning Companions

The preceding analysis of the learning-performance paradox points to a fundamental design problem: the LLMs now entering education were not built for education. They were built for work, and the logic of that design, optimising for output quality, minimising cognitive effort, and treating each interaction as independent and disposable, is precisely what makes them unsuitable as learning tools without deliberate redesign. Understanding what AI for learning must look like therefore requires first being clear about how it differs from AI for work, not just in context but in purpose, design logic, and what it means to succeed.

Figure 1 formalises this contrast across nine dimensions. The most critical distinctions concern purpose, the nature of the interaction, and the relationship to cognitive effort. In work contexts, AI performs or co-performs the cognitive task on behalf of the user, each interaction is transactional and stateless, and success is measured by the quality and efficiency of outputs. In learning contexts, AI must instead scaffold and challenge the learner to produce their own understanding, interactions must be developmental and cumulative, and success is measured by retention, transfer, and metacognitive growth. Equally revealing are the key failure modes: AI for work fails when productivity gains mask skill atrophy over time, while AI for learning fails when task scores improve while knowledge retention declines — the learning-performance paradox made concrete [10, 110]. These failure modes are not accidental but follow directly from design intent. AI for work is designed to minimise friction and provide direct answers; AI for learning must deliberately withhold direct answers and preserve productive struggle as the mechanism through which durable understanding is built [13, 99]. Together these dimensions reveal that the reorientation required is not a minor adjustment to prompting strategy or interface design. It demands a fundamental reconceptualisation of what AI systems in education are designed to do, how they interact with learners, and how their effectiveness is evaluated.

We propose AI learning companions as the design response to this reorientation: adaptive, pedagogically informed, LLM-powered agents designed for sustained integration into diverse educational environments and explicitly engineered to prioritise durable learning over short-term performance. Learning companions comprise three interrelated dimensions that together define how AI can meaningfully support learning as visualised in Figure 2: (1) how students learn with AI, guided by pedagogical principles from the learning sciences; (2) how AI learns about students, through adaptive mechanisms that model and respond to individual progress; and (3) how ethical and human-centred principles ensure these systems remain trustworthy, equitable, and aligned with human values.

The **pedagogical foundation** addresses how students learn with AI. The challenge for AI design is not to reinvent pedagogy but to determine how generative systems can provide the precise support necessary to cultivate both students'

Dimension	AI for Work (Productivity)	AI for Learning (Capability)
Purpose	Complete tasks faster and to a higher standard	Develop the learner's own capabilities over time
What AI Does	AI performs or co-performs the cognitive task	Scaffolds learning and challenges the learner to produce their own understanding
Nature of the Interaction	Transactional and stateless: each interaction is independent	Developmental and cumulative: interactions build on each other over time
Relationship to Cognitive Effort	Reduce effort to maximise efficiency	Preserve and direct effort toward learning
Relationship to Error	Errors are inefficiencies to be eliminated	Errors are diagnostic signals and learning opportunities
Measures of Success	Quality and efficiency of outputs	Retention, transfer, and metacognitive growth
Key Failure Mode	Productivity gains that mask skill atrophy over time	Performance gains that mask absent learning — task scores improve while knowledge retention declines
Design Principle	Minimise friction; direct answers are the goal	Preserve productive struggle; withholding direct answers is a deliberate feature
Theoretical Grounding	Augmentation theory, distributed cognition	Constructivism and self-regulated learning

Figure 1: Comparison of AI designed for work and AI designed for learning, highlighting the reorientation required when moving from professional productivity contexts to educational environments.

cognitive abilities and their metacognitive capacities. To guide this design, we draw on the Higher Education Learning Framework (HELF) [75], which synthesises the learning sciences literature into a set of core pedagogical principles. Building on the HELF, we derive four principles for AI companion design: *deep and interactive learning*, which prompts learners to actively process and connect new knowledge through varied cognitive tasks; *guided scaffolding*, which delivers adaptive guidance calibrated to the learner's Zone of Proximal Development; *learning to learn and higher-order learning*, which fosters metacognitive skills, reflection, and learner autonomy; and *contextual learning*, which situates learning in authentic contexts to shape disciplinary identity and ensure knowledge is transferable. Each principle and its implications for companion design are elaborated in Section 3.

The **adaptive foundation** addresses how AI learns about students. While LLMs offer conversational flexibility and scalability that traditional intelligent tutoring systems could not match, most current LLM applications remain stateless — each interaction begins without memory of the last, making sustained adaptation to individual learners impossible. This foundation resolves that limitation by organising adaptivity into a continuous four-stage cycle: *Capture*, which records learners' digital footprints to gather evidence of engagement and understanding; *Model*, which builds a multi-dimensional learner model from cognitive, affective, and behavioural data; *Adapt*, which determines what, how, and when to adjust learning experiences; and *Evolve*, which evaluates the effectiveness of those adaptations to drive continual improvement. Each stage and its implications for companion design are elaborated in Section 4.

The **responsible design foundation** addresses how AI companions can act with integrity and uphold human values. The highly personalised nature of adaptive AI demands that companions be not only pedagogically effective but also trustworthy, equitable, and aligned with educational values. To guide this design, we organise responsible practice around four interrelated commitments: *security*, which protects learner privacy, data, and digital wellbeing through lawful and responsible data practices; *transparency*, which makes AI processes, reasoning, and data flows visible and explainable to learners and educators; *accountability*, which retains meaningful human oversight and preserves learners' rights to question AI-supported decisions; and *inclusion*, which ensures companions are fair, accessible, and responsive to diverse learners, abilities, languages, and contexts. Each commitment and its implications for practice are elaborated in Section 5.

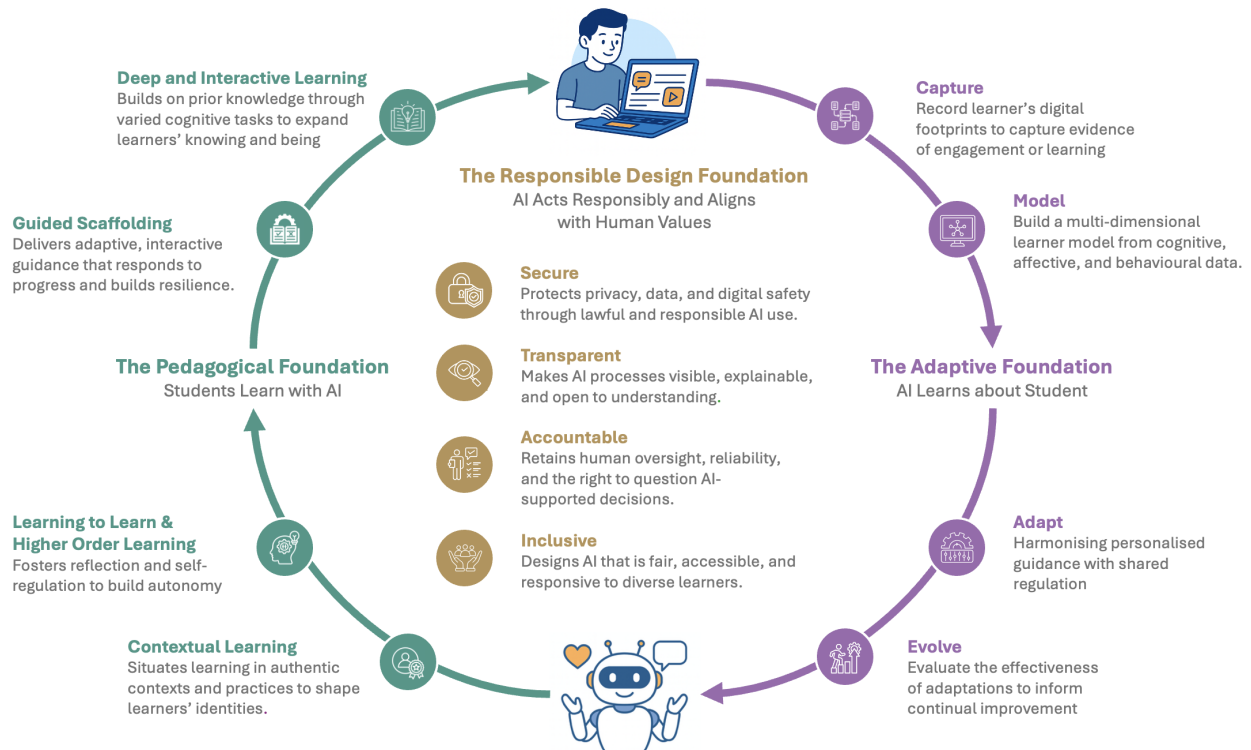


Figure 2: Framework for building AI companions to support learning. The left cycle represents how *students learn with AI* through deep learning, guided scaffolding, metacognitive growth, and contextual engagement. The right cycle represents how *AI learns about students* by capturing, modelling, adapting, and evaluating learning processes. At the centre, *ethical principles*—accountability, transparency, inclusivity, and security—govern the interactions between the two loops.

3 The Pedagogical Foundation: Students Learn with AI

Early applications of LLMs in education have largely relied on prompting strategies that cast the model in roles such as a Socratic tutor (e.g., Khan Academy’s Khanmigo) or study modes offered by OpenAI’s ChatGPT and Google’s Gemini. While useful, these approaches often remain limited in scope and provide only surface-level support. To move beyond this, we draw on evidence from the learning sciences about how people learn best and employ the Higher Education Learning Framework (HELFF) [75] to propose the design of effective AI companions built around four interrelated principles that are essential for fostering robust and equitable learning. These principles are discussed in detail in the following sections.

3.1 Deep and Interactive Learning.

A central goal of education, and one that AI companions must be explicitly designed to support, is deep learning: the kind of durable understanding that transfers across contexts and persists long after instruction ends. Cognitive science provides clear guidance on what produces it. The *generation effect* shows that knowledge constructed by the learner is retained far better than knowledge passively received [92]; the *testing effect* demonstrates that retrieving information from memory strengthens it more than re-reading or re-watching [94]; and research on *desirable difficulties* establishes that conditions which feel effortful in the moment, including retrieval practice, interleaving, and spacing, produce stronger long-term learning than conditions that feel fluent [12]. The practical implication is direct: what matters is not whether a companion presents information well, but whether it makes the student do the thinking. A companion that explains clearly but lets the learner remain passive is, from a learning science perspective, not much better than a well-produced video. Learning is also inherently social. Teaching and learning are grounded in humans’ innate capacity and motivation to connect with others [48], and interaction is not peripheral but integral to meaningful learning. Social dynamics shape motivation, engagement, and higher-order thinking through processes including

action understanding, empathy, and theory of mind [57], the same mechanisms that generate shared understanding and collective engagement in human learning environments [112]. For AI companions, this means that the quality of interaction matters as much as the quality of content: dialogue, perspective-taking, and collaborative sense-making are not add-ons but core mechanisms through which learners construct understanding. *Strategies for AI companions.* To embody these principles, AI companions should be designed to prompt generation and retrieval rather than provision. Rather than explaining a concept, a companion should ask the learner to recall it, apply it, or explain it in their own words. Rather than confirming correctness, it should invite elaboration, surface connections to prior knowledge, and expose gaps in understanding. Companions can also shift fluidly across interactional roles, as tutor, peer, or novice, to sustain dialogue and foster the kind of perspective-taking that amplifies interactive learning. The criterion for good design is not whether the companion feels helpful in the moment but whether it leaves the learner having done more thinking than they would have done alone.

3.2 Guided Scaffolding

A central condition for meaningful learning is that it involves challenge. Cognitive science emphasises that learning is optimised when tasks are neither trivial nor overwhelming, but situated in a “zone of proximal development” where learners can stretch their current abilities with appropriate guidance [108]. Research on desirable difficulties highlights that introducing challenges such as retrieval practice, interleaving, and spacing promotes deeper encoding and transfer, even though these conditions may feel effortful in the moment [92, 12]. Challenge, however, is not synonymous with frustration. Productive struggle occurs when difficulty is balanced with support, enabling learners to persist, adapt strategies, and ultimately experience mastery. Without this balance, excessive difficulty can lead to disengagement, whereas insufficient challenge fosters surface processing and complacency. Thus, challenge and difficulty should be understood as essential mechanisms through which learners refine strategies, consolidate understanding, and develop resilience.

Learning is not solely a cognitive process but is deeply interwoven with emotion. Neuroscience and educational psychology consistently show that emotions influence attention, memory, and motivation [48, 103]. Positive emotions such as curiosity, enjoyment, and hope broaden learners’ willingness to explore and engage, while negative emotions like anxiety or boredom can constrain focus and impede progress. Importantly, negative emotions are not uniformly detrimental: when framed within a supportive environment, experiences of confusion, cognitive dissonance, or even frustration can catalyse deeper inquiry and problem solving [36]. Emotions also serve social and motivational functions, shaping how learners connect with others and how they interpret the value and relevance of their learning experiences. Understanding the role of affect underscores that robust learning environments must not only target knowledge and skills but also cultivate emotional conditions that sustain engagement, resilience, and growth.

Strategies for AI companions. To embody these principles, AI companions should be designed to scaffold productive challenge while also attuning to learners’ emotional states. They can introduce manageable levels of difficulty such as prompting elaboration, posing counterexamples, or encouraging retrieval while offering adaptive support that prevents unproductive frustration. At the same time, AI companions can monitor affective cues in language and interaction, responding with encouragement, reframing confusion as a learning opportunity, or offering strategies for regulation. By balancing challenge with support and integrating sensitivity to emotion, AI companions can help learners experience difficulty as an engine of growth rather than a barrier. In doing so, they contribute to environments where cognitive stretch is paired with emotional scaffolding, enabling learners to persist, adapt, and flourish.

3.3 Learning to Learn and higher order learning

A central condition for sustained educational growth is the capacity to learn how to learn. This involves developing metacognitive awareness: knowing what one knows, monitoring progress, and recognising when strategies need adjustment. Research on self-regulated learning highlights that learners who actively plan, monitor, and evaluate their approaches demonstrate greater autonomy, adaptability, and persistence [120, 85]. Such metacognitive processes are not innate; they must be deliberately cultivated through opportunities for reflection, feedback, and strategy use.

Metacognition is not merely a companion to higher-order thinking but its prerequisite. Meaningful education requires engagement with processes that extend beyond the reproduction of knowledge toward its transformation and application. Bloom’s taxonomy and subsequent frameworks emphasise the importance of analysis, synthesis, evaluation, and creation as core dimensions of higher-order thinking [14, 6], and learners develop deeper understanding when they critique arguments, integrate knowledge across domains, and generate novel ideas. Higher-order learning is also closely linked to transfer, the ability to apply principles in new or unfamiliar settings [84, 16], which requires not only domain knowledge but also the metacognitive awareness to recognise when and how prior learning applies. Together, these

capacities are vital for preparing students to navigate complex, real-world problems where adaptability, creativity, and critical judgment are as important as content mastery.

Strategies for AI companions. To embody these principles, AI companions should be designed to scaffold both metacognitive awareness and higher-order thinking as interrelated capacities. They can prompt learners to articulate their goals, plan strategies, and reflect on progress, thereby strengthening self-regulation and adaptability. A particularly powerful application is metacognitive calibration: comparing a learner's expressed confidence against their actual performance to surface the gap between what they think they know and what they demonstrably know [61]. Students are consistently poor judges of their own understanding, and AI companions are well positioned to make this visible, not by correcting learners directly, but by designing interactions that reveal the limits of their current knowledge and prompt them to adjust their self-assessment accordingly. At the same time, companions can challenge learners with open-ended questions, request justifications for reasoning, and encourage comparisons across contexts to deepen analysis and synthesis. By prompting learners to generate explanations or teach back concepts, AI companions make higher-order thinking processes explicit and accessible, helping learners develop the reflective and critical capacities required for transfer, innovation, and lifelong learning.

3.4 Contextual Learning

Contextual learning emphasises that knowledge is not acquired in isolation but is shaped by the environment, tools, and practices in which it is used. Situated learning theory highlights that authentic contexts provide learners with opportunities to engage in tasks that mirror real-world practices, enabling knowledge to be both applied and understood in relation to its use [17]. Kolb's experiential learning cycle highlights the iterative interplay of experience, reflection, conceptualisation, and experimentation [60], while Shulman's signature pedagogies and the theory of threshold concepts emphasise the transformative practices that shape professional identity [97, 69].

Learning is also a process of becoming. From this perspective, education is not simply about acquiring skills or information but about shaping identities, dispositions, and ways of participating in communities of practice [111]. Becoming a learner in a discipline involves appropriating its language, methods, and values while negotiating one's own sense of purpose and belonging. Such identity work is developmental, as students shift from peripheral participation to more central roles in collaborative practices. This view underscores that learning is relational and forward-looking, oriented toward the kinds of person learners are becoming and the roles they aspire to inhabit in society [9].

Strategies for AI companions. To embody these principles, AI companions should be designed to situate learning in authentic contexts and scaffold the development of professional identity. They can frame problems through real-world scenarios, case studies, or discipline-specific practices, helping learners connect abstract concepts to concrete applications. Building on experiential learning cycles, AI companions can prompt learners to reflect on prior experiences, test new ideas, and refine understanding through iterative experimentation. They can also model or simulate aspects of signature pedagogies so that learners practice the distinctive ways of thinking and acting valued within their field. Finally, AI companions can support learners in navigating threshold concepts by identifying sticking points, reframing misunderstandings, and encouraging persistence through transformative moments of learning.

While these principles guide how students learn with AI companions, effective partnership also requires that the AI itself learns about the student — constructing models, adapting behaviour, and evolving over time.

4 The Adaptive Foundation: The AI Learns about the Student

For learning companions to support learners effectively, they must themselves be capable of learning about the learner. This adaptive dimension concerns how AI captures, models, and interprets learner data to tailor guidance, feedback, and sequencing, and how it evolves through continual evaluation. Building on four decades of research in adaptive and intelligent tutoring systems, this section outlines a four-stage adaptive cycle, *Capture, Model, Adapt, and Evolve*, that enables AI companions to transform raw learner interactions into personalised and pedagogically grounded support.

4.1 Capture: Recording Learners' Digital Footprints

A defining advantage of digital learning environments is their ability to record nearly every learner action, creating detailed logs of clicks, keystrokes, problem-solving steps, time spent, and communication. Decades of research in learning analytics and educational data mining have established that these fine-grained digital traces can serve as meaningful proxies for underlying learning processes [80]. The broader challenge, of course, lies in making sense of this deluge of raw data. There is often a gap between low-level interaction data and the high-level constructs they reflect.

In other words, while we can capture what learners do, interpreting those actions in terms of learning gains, strategies, or misconceptions requires careful modelling and remains an ongoing focus of research [87].

This challenge is multiplied by advances in sensing and multimodal interfaces that now allow the capture of an even richer array of learner signals. *Multimodal learning analytics* (MMLA) integrates data from sources such as eye-tracking, gaze, facial expression, speech, and physiological sensors (e.g., heart rate) with traditional clickstream logs. By combining these channels and the use of AI, researchers can form a more holistic picture of learning, linking observable behaviours with internal cognitive and affective states.

The addition of LLMs introduces significant new affordances in this capture phase [56, 64]. Unlike legacy platforms that log only discrete events, GenAI-enabled environments capture rich, open-ended learner–AI dialogues, including prompts, explanations, reflections, and feedback. These interactions form a new class of trace data—naturalistic, conversational, and embedded in process—which can reveal student epistemic frames, strategic decisions, and cognitive engagement in ways traditional logs cannot [62]. Furthermore, GenAI systems themselves generate metadata—rationale, summaries, uncertainty scores—that can be incorporated into LA pipelines to enrich learner models and support interpretability.

Strategies for AI companions. Implementing the capture phase with AI companions involves integrating traditional data collection methods with new generative capabilities. Learning management systems and educational data mining techniques reliably log structured data—quiz scores, time on task, or resource views—while LLMs help interpret unstructured learner outputs, such as free-response answers or forum posts, extracting key themes, misconceptions, or even sentiment (e.g., frustration or confidence) [50]. Machine learning classifiers can monitor behavioural signals like idle time or erratic clicking, while AI with speech or vision capabilities can infer affective states from tone or facial expression during video-based learning. These traces should be collected intentionally and ethically in line with established privacy and analytics frameworks [38]. Together, these signals form the epistemic foundation of the dual-loop framework: they allow the companion to personalise support, foster reflection, and evolve with the learner. In this sense, capture is not a passive logging activity but an active design layer that shapes how adaptivity is realised in practice.

4.2 Model: Building a Multi-Dimensional Learner Model

Once data are captured, the system’s next task is to model the learner, constructing a representation that encapsulates the student’s current state of knowledge, skills, strategies, and affective dispositions. Research in student modelling since the 1990s has explored how learner data can be used to drive intelligent adaptation in learning systems [107]. This work has yielded models capable of detecting a wide range of learner dimensions: which skills students know [82], how long their memory of something learned will last [81], what self-regulated learning strategies they employ [72], and whether they are engaged or disengaged at any given moment [7]. Although these models were rarely implemented at scale before the emergence of large language models (LLMs), with the notable exception of knowledge models used to decide when to advance the student [59]—they provide a strong foundation for informing the decisions of modern AI systems.

A key example is *knowledge tracing* [26], which encodes what a student currently knows relative to a domain model, a structured representation of the concepts, skills, or knowledge components that define a subject. While LLMs can be prompted or fine-tuned to diagnose specific student errors or provide feedback for incorrect solutions [100], they are less effective at inferring a student’s overall level of mastery for a skill. This is because such inference requires integrating data across all prior interactions, weighting recency and frequency, and estimating the probability of mastery [42]. Fortunately, algorithms already exist to perform these probabilistic computations, and hybrid approaches that combine LLMs (for content interpretation and response analysis) with traditional knowledge-tracing algorithms (for longitudinal integration) have been shown to outperform either method alone [95]. A similar synergy applies to memory algorithms, which can predict when a student will forget previously learned material and identify optimal review intervals [81].

Decades of research on other aspects of learner modelling—such as strategies, engagement, and emotion—have also produced algorithms that can provide actionable information to LLMs. The LLM does not need to replicate these specialised functions; instead, outputs from such models can be fed into the LLM as structured text or numeric representations. In some cases, such as analysing student writing or explanations, the LLM may outperform earlier approaches, while in others—such as evaluating behavioural logs or facial expressions—dedicated machine learning models remain superior. The most effective approach combines the interpretive strength of LLMs with the precision of domain-specific models, producing a richer, more comprehensive picture of the learner.

Despite these advances, previous generations of ITSs often struggled to leverage student models for more than a single adaptive purpose within one system [8]. The most common application was mastery determination, deciding whether a student should move to the next topic [59]—while others focused on topic sequencing [28] or affective support [116]. This limitation likely stemmed from the human design effort required to coordinate multiple adaptive dimensions.

AutoTutor, one of the most comprehensive ITSs, explored many data-driven forms of adaptivity [76], yet it required decades of collaboration across dozens of researchers and typically studied one form of adaptation at a time. Emerging AI systems can now reason about unique combinations of learner information, experiment with alternative instructional strategies, and update their internal models based on outcomes. Just as Reinforcement Learning from Human Feedback (RLHF) has driven the development of general-purpose chatbots, *Reinforcement Learning from Human Learning* (RLHL) [37] may enable the next generation of generative AI-powered tutoring systems—ones that continually learn from interactions with students to refine their own adaptive policies.

Strategies for AI companions. Building on these foundations, AI companions should adopt hybrid modelling architectures that integrate the interpretive capabilities of LLMs with established learner-modelling algorithms. LLMs can analyse open-ended learner responses to extract conceptual misunderstandings, affective cues, and evidence of metacognitive behaviour, while probabilistic or mathematical models such as knowledge tracing, memory modelling, and engagement detection integrate these signals across time. This fusion supports multi-dimensional learner models that evolve dynamically with the learner. To enhance transparency and trust, AI companions should make these models visible and interpretable through open learner model interfaces, allowing both learners and instructors to inspect, question, and adjust the AI’s inferences. Over time, such systems can employ reinforcement learning approaches to optimise how learner data are used to personalise feedback and adapt instruction, creating companions that learn not only about the student but also from them, continually refining their pedagogical strategies to maximise growth, equity, and understanding.

4.3 Adapt: Harmonising Personalised Guidance with Shared Regulation

Adaptive and intelligent tutoring systems are designed to modify their pedagogical decisions dynamically in response to individual learner characteristics (e.g., prior knowledge, affect, and engagement) and real-time learner actions [4]. A classical way to conceptualise adaptivity is through two interrelated “loops” [105]. The *outer loop* operates at the macro level, managing the selection and sequencing of learning tasks. Its main responsibility is to determine which activity or problem the learner should tackle next, effectively tailoring the curriculum pathway to the individual’s progress and mastery level. The *inner loop* functions at the micro level within a single task, providing step-by-step support as the student works through a problem. This includes checking correctness, offering error-specific feedback, providing hints, and scaffolding problem-solving processes [4, 26]. Together, these two layers enable an ITS to adapt both *what* content a learner encounters (macro-adaptation) and *how* feedback and guidance are provided during task execution (micro-adaptation). For example, an outer loop might decide to reteach a struggling student a particular concept using an easier problem, while the inner loop on that problem provides immediate hints and feedback as the student attempts it.

A substantial body of evidence has demonstrated the effectiveness of adaptive learning systems relative to both classroom instruction and non-adaptive educational technologies [5, 106, 67]. These systems improve cognitive efficiency and learning outcomes, yet this automation often comes at a cost. When the system assumes full control of pacing, sequencing, and support, students may become passive recipients of guidance, offloading metacognitive control and losing opportunities to develop self-regulation [71]. In fully automated tutors, learners have little reason to set their own goals, decide which topics to review, monitor their understanding, or plan their study, as the system performs these functions on their behalf.

To counteract this limitation, recent research has shifted toward designs that integrate principles of socially shared regulation of learning [79] and hybrid human–AI co-regulation [71]. Rather than fully automating adaptivity, these systems invite learners to share control over decision-making. Transparency mechanisms, such as Open Learner Models (OLMs), make the system’s inferences visible, prompting learners to reflect on their understanding and progress [20]. Similarly, adaptive systems can embed prompts that involve learners in regulatory choices, such as asking whether they would like a hint or prefer to continue independently [73]. Such participatory features engage learners in monitoring and control, preserving the efficiency of adaptivity while nurturing metacognitive growth [2]. Goal setting plays a crucial role in this process: it anchors monitoring and self-assessment, ensuring that learners can evaluate progress against explicit objectives [78, 115]. Without clear goals, students often struggle to activate their internal “monitoring and control” loops effectively [114].

Strategies for AI companions. The rise of LLMs introduces powerful new affordances to extend and humanise adaptivity within learning companions. Traditional adaptive systems relied on a fixed pool of pre-authored problems, hints, and explanations, constraining flexibility and responsiveness. In contrast, generative AI enables dynamic co-construction of learning activities, allowing the outer loop to operate interactively in dialogue with the student. Learners’ curiosity, goals, or misconceptions can directly influence the next problems or explanations generated. Another key affordance is explainability: generative AI tutors can articulate the rationale behind their instructional decisions in natural language. Whereas earlier adaptive systems could often not articulate why they evaluated a student in a certain way, LLM-based

companions can express reasoning such as, “Your quick reading of high-level help resources and subsequent correct answer, combined with past correctness, suggests that your error was based on forgetting a function name rather than something deeper. This transparency fosters trust and self-reflection, aligning with research on learning analytics dashboards and Open Learner Models that visualise learners’ knowledge states to promote metacognitive awareness and agency [20]. Beyond transparency, generative AI companions can actively scaffold self-regulated learning through reflective dialogue—asking, for example, “What is your goal for this session?”, “How confident do you feel about this topic?”, or “Can you explain how you solved that problem?”. Such conversational prompts have long been recognised as effective in eliciting planning, monitoring, and self-assessment behaviours. Through these exchanges, the AI companion not only adapts instruction but also cultivates the learner’s capacity to themselves adapt. This represents a significant philosophical shift: from adaptivity being something *done to* the learner, where the system unilaterally controls pacing and feedback, toward adaptivity *done with* the learner, in which AI acts as a co-regulator and partner in metacognitive growth, balancing personalisation with shared regulation.

4.4 Evolve: Continuous Improvement Through Feedback Loops

The final stage of the adaptive cycle is to evolve the system itself: evaluating how well its adaptations are working and refining them for continual improvement. Essentially, the system uses feedback not only to help the student but also to help itself, learning from its own successes and failures. In the context of adaptive learning, this concept has been formalised as *design-loop adaptivity* or *closed-loop evaluation*, where learner data are used to iteratively redesign and improve the instructional system [4].

Without this evaluative phase, adaptive systems risk stagnation: their models and interventions may remain static despite shifts in curricula, contexts, or the learners themselves. The evolve phase therefore completes the adaptive cycle by treating the AI as a learning entity—one that continuously refines its instructional policies based on empirical evidence rather than fixed design assumptions. This adaptive meta-learning ensures that the system’s pedagogical foundations remain robust, equitable, and contextually responsive. In other words, the AI not only models the learner but also optimises how it supports learning.

Strategies for AI companions. To operationalise the evolve phase, AI companions should treat instructional decisions as hypotheses to be tested rather than fixed policies to be applied. At the system level, controlled experiments and multi-armed bandit algorithms can dynamically allocate learners to instructional conditions shown to be more effective, improving both learner outcomes and system understanding over time [88]. When a pedagogical strategy consistently fails to improve performance, this serves as a diagnostic signal to modify or replace it, an iterative process analogous to formative evaluation in education. Larger-scale data enable the system to go beyond learning what works to learning what works for whom and under which circumstances, supporting equitable as well as effective adaptation. Simulated learners that model realistic student behaviours offer a complementary approach, allowing new instructional strategies to be tested rapidly before deployment [41]. A more recent innovation leverages generative AI as a co-designer: LLMs can propose new feedback prompts, scaffolding sequences, or dialogue structures that are then validated through empirical testing with real learners. Together, these strategies ensure that the AI companion does not merely adapt to the learner in the moment but continuously improves how it supports learning over time.

5 The Responsible Design Foundation: AI Acts with Integrity and Upholds Human Values

The highly personalised and embedded nature of learning companions means that their influence extends well beyond technical optimisation. As companions capture rich learner data, shape study habits, and mediate everyday decisions about effort, strategy, and assessment, they also carry social, ethical, and institutional consequences. Responsible design is therefore not a peripheral concern but a core requirement: learning companions must protect learners from harm, make their operations intelligible, remain under meaningful human oversight, and promote equity rather than reproduce existing disparities.

Trustworthiness is the unifying goal of this foundation, encompassing the expectations of students, educators, technologists, institutional leaders, and where relevant, parents and caregivers. We organise this around four interwoven commitments: *security*, *transparency*, *accountability*, and *inclusion*. While these commitments apply to any AI application, we clarify how they play out distinctively with learning companions, and the specific choices that platform owners must make about the data they capture, how it is used, and by whom.

5.1 Security: Protects privacy, data, and digital safety through lawful and responsible AI use

UNESCO guidance emphasises that AI systems must be mindful of data ownership, data privacy, and data availability for the public good, and should adopt principles of privacy and security by design [70]. Without robust safeguards,

the rich interaction data generated by learning companions could be misused for unintended profiling, surveillance, or exposure of private information. Secure practice includes separating personally identifiable information from interaction data, preferring local or institutionally hosted processing for sensitive tasks, and implementing tiered retention with auditable deletion logs.

The interaction logs and transcripts generated by AI companions can range from relatively innocuous content, such as problem-solving traces in an algebra tutor, to sensitive personal disclosures, such as emotional reflections on a difficult work placement shared with a conversational companion. Whether educational institutions are licensing commercial products hosted remotely, deploying on-premises, or adopting a hybrid approach, they must make deliberate choices about how much data they archive and for what purposes.

Security is therefore not only technical but extends to psychological safety. Given the social and conversational nature of AI companions, learners may disclose more than they intend or become emotionally dependent on interactions in ways that require careful design. Companions should protect learners' digital wellbeing by avoiding manipulative designs, implementing age-appropriate guardrails, and preventing harmful or crisis-related content from bypassing safety layers. Robust defences against prompt injection, content exfiltration, and adversarial misuse are essential, as are proactive risk-scanning mechanisms that detect anomalous system behaviour. By treating data protection and wellbeing as preconditions for adaptivity rather than trade-offs, security becomes an enabler of trustworthy personalisation.

Strategies for AI companions. To operationalise security and safety, learning companions should employ modular architectures that isolate sensitive processes, adopt privacy-preserving representations such as derived or aggregated features rather than raw traces, and automate privacy notifications so learners understand what is stored and why. Synthetic or differentially private data can be used for model updates where possible. Periodic self-checks, anomaly detection pipelines, and red-team testing ensure secure operation across the learning lifecycle. Companions should align their prompts, retrieval, and sensitive content filtering with institutional academic-integrity settings, and maintain stable, versioned model deployments so that changes do not invalidate the integrity of established companion behaviour.

5.2 Transparent: Makes AI processes visible, explainable, and open to understanding

The literature consistently underscores that transparency and explainability are not optional add-ons but foundational requirements for AI-powered educational tools. These requirements stem from two interrelated imperatives: establishing calibrated trust and enabling meaningful, pedagogically actionable insight for all stakeholders—students, educators, and institutional decision-makers. When AI systems make decisions that influence learning pathways, assessment outcomes, or resource recommendations, opaque processes undermine confidence. Learners may over-trust fluent but unjustified suggestions, while educators may resist adoption altogether when they cannot see why or how the system arrived at its outputs. This tension is especially acute for complex “black-box” architectures, such as deep neural networks, whose internal logic is unintelligible without deliberate design for interpretability.

Explainable AI (XAI) frameworks address this challenge by making the system's reasoning visible and comprehensible. In education, however, explanation must be pedagogically purposeful, not merely technically informative. The XAI-ED framework [53] formalises this insight, arguing that explanations should actively advance teaching and learning—for instance, by helping students understand misconceptions or by enabling educators to diagnose why a learner model reached a particular conclusion. XAI thus supports fairness, accountability, and ethical integrity in high-stakes uses of AI—such as automated grading, risk prediction, and personalised feedback—mitigating concerns about bias and undue influence [47]. Ultimately, transparency and explainability ensure that AI systems operate as accountable partners within the learning ecosystem.

In the case of AI tutors, transparency and explainability can also build directly on a long-standing tradition of Open Learner Models (OLMs) [21], which make the system's representation of a learner's knowledge, skills, or progress available for inspection and negotiation. Classic OLM work demonstrates that simply exposing the model fosters trust, supports metacognition, and encourages learners to take a more active role in monitoring and directing their progress. Integrating XAI enables OLMs to reveal the evidential basis for inferences (e.g., “This misconception was detected because your last three attempts showed. . .”) and the rationale behind pedagogical decisions (e.g., “This resource is recommended because it addresses the specific pattern of errors you exhibited in. . .”). This transformation reframes the OLM from a passive display into an active pedagogical instrument. With transparent, explanation-rich learner models, students can interrogate AI recommendations, challenge inaccuracies, and understand the logic behind their personalised pathways.

Strategies for AI companions. To build calibrated trust and enable meaningful learner agency, learning companions must adopt a suite of strategies that foreground transparency in both their operations and their pedagogical interactions. First, companions should clearly communicate what data they collect, how it is used, and which aspects of the learner model inform specific recommendations or feedback. This includes providing layered explanations that range from

simple justifications (e.g., “I recommended this resource because. . .”) to more detailed technical rationales for educators or advanced users. Second, companions should make their reasoning processes visible by revealing the evidence, inference steps, or patterns that shaped their conclusions, allowing learners and teachers to interrogate, contest, or refine the model’s understanding. Third, companions should openly disclose uncertainties and limitations—highlighting when an inference may be incomplete or when data is insufficient—to prevent overconfidence and encourage critical engagement. Fourth, transparent companions should make their capabilities, constraints, and guardrails explicit, helping learners calibrate expectations and avoid inappropriate reliance on AI-generated suggestions. Finally, transparency must extend to the evolution of the learner model over time, showing how new actions or responses update the system’s interpretation of progress.

5.3 Accountable: Retains human oversight, reliability, and the right to question AI-supported decisions

Accountability ensures that companions remain corrigible, governable, and aligned with educational responsibility. Even when models are highly capable, educators and institutions, not algorithms, retain responsibility for learning experiences and outcomes. This requires meaningful human oversight, reliable performance, and clear avenues for contesting AI-supported decisions [113, 40].

High-stakes inferences, including mastery promotion, at-risk flags, and achievement predictions, should incorporate human-in-the-loop checkpoints. Learners and teachers require channels to dispute model states or recommendations, and these disputes should feed into refinement pipelines or policy-rule updates. Reliability depends on continuous monitoring of hallucination rates, harmful suggestions, and model drift, complemented by retrieval-augmented generation, domain-validated templates, and specialised models such as knowledge tracing that constrain LLM free-form generation. Audit logs linking inputs, model states, actions, and outcomes enable rigorous incident analysis.

Crucially, accountability must be learning-aligned. Companions should be evaluated not only on output accuracy but on their impact on retention, transfer, metacognition, and durable learning [13, 99]. This connects directly to the paper’s central argument: a companion that produces high task performance while undermining genuine learning has failed its core purpose, regardless of how technically reliable it appears. Ensuring that adaptivity enhances rather than bypasses productive struggle is therefore a core accountability criterion, not a secondary consideration.

Strategies for AI companions. Companions can include review modes for educator verification of system actions, implement dispute workflows for learners, and maintain auditable decision trails. They should run continuous reliability benchmarks and use fallback behaviours, such as simplified rule-based responses, when confidence drops. Governance reviews and incident post-mortems ensure that companions remain reliable, correctable, and aligned with educational goals. Evaluation frameworks should incorporate delayed retention and transfer measures rather than relying solely on immediate performance metrics.

5.4 Inclusive: Designs AI that is fair, accessible, and responsive to diverse learners

Inclusion requires that learning companions expand opportunity rather than personalising disadvantage. Because models are trained on data reflecting historical inequities, they risk reproducing bias in content, tone, and adaptive decisions. Inclusive companions therefore embed algorithmic fairness, accessibility, and cultural responsiveness at every stage of design and deployment.

Datasets and outputs should be audited for representational harm and performance disparities across demographic groups [11, 22, 89]. Fairness-aware training and post-processing can mitigate some risks, but inclusive design must also attend to interactional accessibility and contextual appropriateness. Following Universal Design for Learning principles, companions should support multiple means of engagement, representation, and expression—adjustable reading levels, multimodal explanations, captioning, multilingual support, alternative input modes [68].

Supporting low-bandwidth and mobile-first environments is essential to prevent digital exclusion. Participatory co-design with under-represented learners and educators ensures that prompts, examples, scaffolds, and help policies reflect local curricula and cultural norms [19]. Finally, inclusive companions must counter risks of cognitive offloading by providing staged hints, self-explanation prompts, and faded support so that help develops learner autonomy rather than dependency [91].

Strategies for AI companions. Inclusive companions can dynamically adjust language complexity, offer multilingual and multimodal options, and apply fairness-aware personalisation that tracks disparate effects. Accessibility-first modes can prioritise simplified layouts, alternative inputs, and sensory-friendly feedback. Co-design cycles with diverse learners help ensure grounding in real contexts. Fairness dashboards, usability audits, and localisation reviews maintain inclusivity as systems evolve.

6 Case Studies of LLM-Powered Tools Designed to Support Learning

This section presents five case studies of LLM-powered educational systems intentionally designed to support student learning across diverse contexts and disciplines. Each case study is authored by researchers who led or made significant contributions to the development of the tool described, providing firsthand insight into the design decisions, empirical findings, and practical challenges involved.

Section 6.1 presents Khanmigo, an AI-powered tutor embedded within the Khan Academy platform, designed to scaffold skill practice through guided hints, worked examples, and cognitive engagement measurement across a broad global curriculum. Section 6.2 presents RiPPLE, a university-scale learning platform that integrates generative AI companions into cycles of student content creation, peer evaluation, and personalised practice, supporting active and social learning at scale. Section 6.3 presents CodeHelp, which provides scalable, solution-free programming support through structured queries and guided tutoring workflows, modelling productive struggle and effective help-seeking as a metacognitive skill. Section 6.4 presents JeepyTA, an AI teaching assistant embedded in course discussion and feedback workflows, supporting dialogue, formative feedback, and brainstorming while preserving configurable instructor oversight. Finally, Section 6.5 presents Recast, an institutional platform developed at the University of Technology Sydney that enables educators to design and deploy steerable, context-sensitive AI assistants grounded in course-specific resources, with full enterprise governance and deliberate attention to responsible design at scale. For each tool, we describe its purpose and deployment setting and analyse how its design aligns with the pedagogical, adaptive, and responsible foundations articulated earlier in the paper. Collectively, these cases illustrate how generative AI can move beyond task-oriented assistance toward learner-centred support that promotes understanding, reflection, and sustained engagement across a wide range of educational contexts.

6.1 Khanmigo: An AI-Powered Tutor Integrated into a Global Learning Platform

6.1.1 Overview of Khanmigo

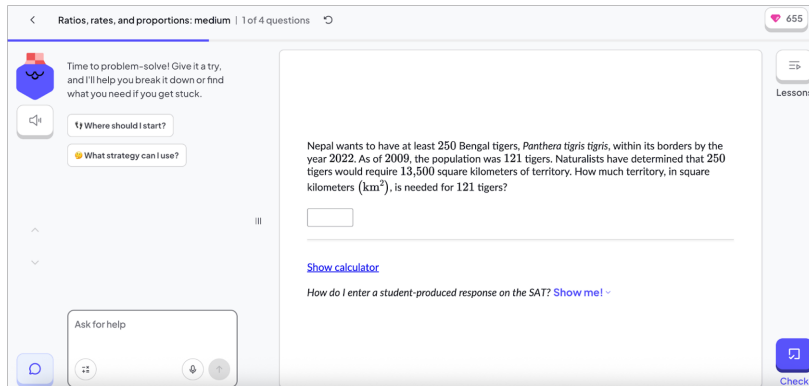
Khan Academy is a supplemental online instruction and practice platform meant to provide students with the independent practice opportunities they often do not receive with their core curriculum. Khanmigo is an AI-powered tutor integrated into this practice platform to provide scaffolding and feedback in ways not available prior to the advent of generative AI. In the 2025-26 school year, approximately 1M students and teachers will engage with Khanmigo in classrooms as part of the Khan Academy Districts offering. Research across multiple datasets from multiple digital learning platforms has demonstrated that each practice opportunity for students results in predictable gains toward skill mastery [58]. Unfortunately, students do not consistently get the opportunity to engage in the amount of practice needed to reach mastery. Khan Academy and Khanmigo are meant to change that.

Learners interact with Khanmigo in a number of ways. First, Khanmigo is available for students as they are working on Khan Academy content (see Figure 3a), in what Khan Academy calls companion mode (not to be confused with the term companion used elsewhere in this paper). Second, students can bring outside questions and problems to Khanmigo in TutorMe mode (see Figure 3b). Third, students can interact with Khanmigo in the essay coach feature, where Khanmigo provides feedback and coaching to students on their writing assignments. Finally, there are 20+ other distinct activities where learners interact with Khanmigo in specific ways, for example in a debate, to play a vocabulary game, or in co-creating a story. For the purposes of simplicity, we will not address the latter category in this paper.

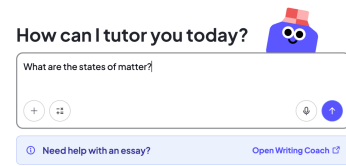
6.1.2 Pedagogical Foundations of Khanmigo

Khanmigo is not designed to be a complete learning companion. As such, it does not meet all of the pedagogical foundations described above and is not intended to. Rather, it is designed explicitly to do what is needed to enhance the practice in Khan Academy's system, mainly by providing improved scaffolding and feedback during practice.

Focus on student cognitive engagement. The initial efforts to design Khanmigo focused on the tutor moves it would make, drawing on literature about what good human tutors do (e.g., [43]). It became clear that Khanmigo could in fact replicate good tutor moves. However, the real problem to tackle was student cognitive engagement. Examination of transcripts revealed too many students simply typing “I don’t know,” “idk,” and other generally passive responses. In order to address this issue, there was a need for a way to measure the cognitive engagement of students. Chi and Wylie’s (year?) ICAP framework provided a helpful taxonomy by which students’ chats with Khanmigo could be categorized as passive, active, or constructive [109]. An LLM judge was created to do this categorization automatically. Initially, the findings of the categorization were disappointing. Students were passively engaged at levels that meant Khanmigo was not likely to improve learning outcomes for them. However, having the measure meant the Khan Academy team could test the impact of changes, like those described below, on engagement, and are now able to adjust Khanmigo behavior to optimize for student cognitive engagement.



(a) Khanmigo in companion mode working with a Khan Academy question.



(b) Khanmigo in TutorMe mode working on a question the student inputs.

Figure 3: Examples of Khanmigo operating in different interaction modes. The left image shows companion mode, in which Khanmigo supports a student working through a Khan Academy problem with hints and guided reflection. The right image shows TutorMe mode, in which students bring their own questions and problems to the system

The walkback of socratic tutoring. The very first prompt ever made to a pre-release version of GPT-4 by Khan Academy staff started with the direction, “You are a socratic tutor.” Much early effort in prompt engineering went into preventing the tutor from giving answers, and creating evaluations to monitor whether it did. However, review of transcripts of students interacting with Khanmigo and a re-read of research literature led to a change in approach. Transcripts revealed significant frustration in students who legitimately did not know content when Khanmigo continued to just ask question after question. This led to the abandonment of interaction with Khanmigo rather than learning. Reviewing the intelligent tutoring literature revealed studies that seemed to suggest that when faced with a problem, if a student has any idea how to solve it, they should make an attempt. If they do not, they should get a hint [3]. Note that a hint is different from a socratic question. In response, the team redesigned Khanmigo in companion mode so that it encourages students to make an attempt and provides hints when needed. Then, if the learner’s attempt is incorrect, Khanmigo engages deeply with the student on the problem, encouraging reflection on where they went wrong, including providing the answer in the form of a worked example.

6.1.3 Adaptive Foundations of Khanmigo

The capture and modeling of student proficiency has long been a feature of Khan Academy (prior to the introduction of Khanmigo). The modeling of student proficiency is done at a much more simplistic level than modeling in many intelligent tutors. Students meet clear heuristic rules to move their status on each skill from attempted to familiar to proficient to mastered. For example, when working on an exercise, getting 100% results in proficient status. Getting questions on that skill correct on a mixed practice assignment, like a unit test or course challenge, bumps that status to mastered. Students are able to see their current status on every skill in their course (or other courses they wish to explore). At one time, Khan Academy had a more probabilistic algorithm working behind the scenes to both estimate mastery and make recommendations. Students vociferously shared that they did not understand what they needed to do to get to mastery; they wanted to have a clear understanding of how many questions on what assignments they needed to get right to move up. This is understandable if their homework assignment is “get to mastery on systems of equations” rather than “complete all the questions in the exercise” which is typical in many curricula. In the same vein, teachers did not understand why the system was making particular recommendations, and many of the recommendations did not align with the curriculum being taught in class. So, the system operates on clear rules and students are assigned groups of skills in units on which to get mastery, often by their teacher rather than the system.

The mastery system in Khan Academy provides a basis for adaptivity in conversations with Khanmigo. When a student is working on a problem with Khanmigo in companion mode, the system inserts the students’ mastery level on that skill and pre-requisite skills into the prompt. This has shown promising results in improving student success on future problems on that skill. Khan Academy also experimented with collecting student interests and injecting them into prompts to influence conversations but has not resulted in improvements to student engagement or learning to date. For all the interest in personalization, so far Khanmigo has been most successful with information about how much the student knows about the current skill being practiced.

6.1.4 Responsible Design Foundations of Khanmigo

Given Khanmigo is largely deployed with learners in U.S. grades 3-12 in schools, significant attention was paid to safety and security.

Moderation and transparency of student conversations. Khanmigo involves minors interacting with LLMs, so there was a need for visibility into their conversations. For students in classrooms, their entire conversations are viewable by their teacher (and conversations in individual accounts are visible to parents). Student responses are all sent through a moderation AI that flags instances of violence, hate, self-harm, sexual content, and other concerning issues. If a response meets the criterion, it is flagged to the teacher or parent and Khanmigo redirects the conversation. There is a trade-off in the system between catching all concerning interactions and false positives. Some conversations, like discussions of the deaths in *Romeo & Juliet*, are generally going to cause flags. The balance is to make sure that teachers are not receiving so many flags that they begin to ignore them.

Data privacy and security. Khan Academy maintains agreements with the foundational model providers it uses stating that no user data will be used to train their models. The agreements also contain clear data deletion policies where data are only held by the model providers long enough to ensure there are no concerns about safety, and then they are deleted. Similarly, Khan Academy has data retention policies that outline when chats will be deleted, given data deletion is the best guarantee against breaches.

6.2 RiPPLE: An AI-Powered Student Co-Creation and Peer Learning Platform

6.2.1 Overview of RiPPLE

RiPPLE is an AI-enhanced learning platform designed to support active, social, and personalised learning at scale. Used by more than 80,000 students across multiple disciplines, RiPPLE situates learners at the centre of a triadic cycle of creation, evaluation, and practice, supported by peers, educators, and, more recently, AI learning companions [55]. Figure 4 illustrates this ecosystem and the platform’s evolving integration of generative AI.

Panel (a) depicts RiPPLE’s core workflow. Students generate bite-sized learning resources—multiple-choice questions, explanations, worked examples, or conceptual problems—functioning as “experts-in-training”. Because student-generated resources may include inaccuracies or incomplete pedagogical reasoning, each resource is evaluated by multiple peers and, where necessary, by instructors. High-quality resources are published to a shared repository; those requiring improvement are returned to authors for revision. Across these activities, RiPPLE continuously infers learners’ topic-level abilities and recommends personalised practice resources aligned with their evolving mastery.

Building on this foundation, RiPPLE now integrates a suite of generative AI learning companions that support students across all three learning activities. Panels (b), (c), and (d) illustrate how companions provide real-time guidance during content creation, scaffold constructive peer evaluation, and enhance personalised practice through adaptive explanations, just-in-time prompts, and targeted follow-up questions. These capabilities extend RiPPLE’s capacity to provide scalable, context-aware support while preserving learner agency and pedagogical integrity.

6.2.2 Pedagogical Foundations of RiPPLE

RiPPLE’s design is grounded in well-established principles from the learning sciences that emphasise active knowledge construction, generative engagement, and iterative cycles of practice. Rather than positioning learners as passive recipients of information, the platform structures learning around cognitively demanding tasks that foster elaboration, retrieval, metacognition, and peer-supported reasoning. The integration of generative AI learning companions is intentionally aligned with these mechanisms: companions do not simply provide answers, but serve as prompts, scaffolds, and reflective partners that help learners think more deeply, regulate their learning, and engage productively with challenging material. The following sections outline how RiPPLE operationalises these pedagogical foundations across creation, peer evaluation, and personalised practice.

Elaboration through content creation and peer feedback. A defining pedagogical feature of RiPPLE is its requirement for students to generate learning resources, a process shown to enhance student engagement, promote deeper processing, and support meaningful learning [54]. The cognitive effort required to explain a concept, select core ideas, anticipate misconceptions, and craft plausible distractors promotes elaboration, a central mechanism for deep learning [45]. Structured peer feedback amplifies these benefits. When reviewing peers’ contributions, learners must assess correctness, clarity, and pedagogical value, requiring them to articulate reasoning, detect conceptual gaps, and consider alternative perspectives [102]. This reciprocal process of creating and evaluating resources positions learners simultaneously as teachers and critics, strengthening their elaboration and metacognitive awareness. AI learning companions enhance both creation and peer evaluation by prompting deeper cognitive engagement. During creation,

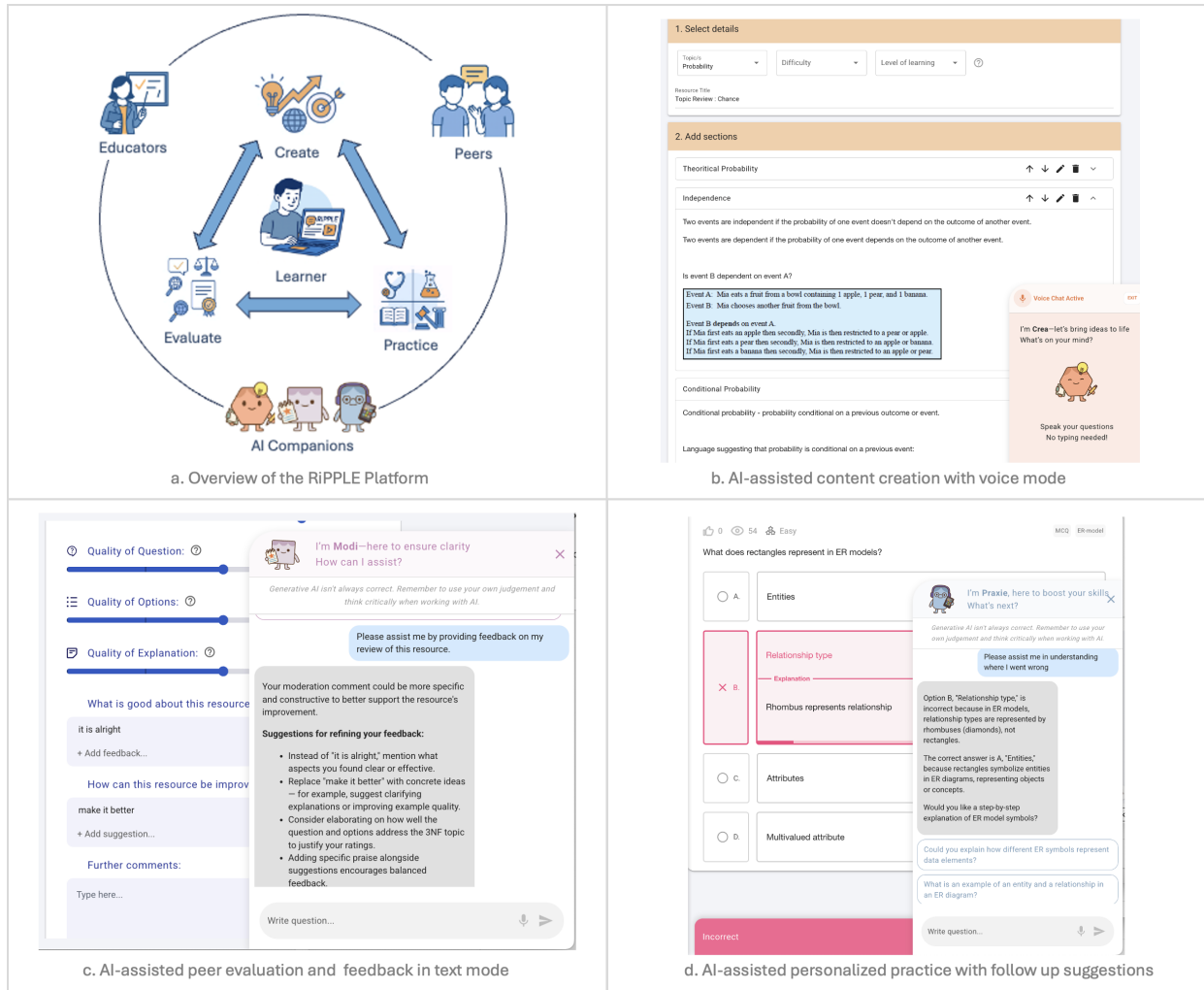


Figure 4: Illustrative overview of the RiPPL platform and its AI-assisted learning workflows. Panel (a) presents the core learning cycle—creation, evaluation, and personalised practice—supported by educators, peers, and AI companions. Panel (b) shows AI-assisted content creation with optional voice-based interaction. Panel (c) demonstrates AI-supported peer evaluation, where learners receive constructive guidance to refine their moderation comments. Panel (d) depicts AI-enhanced personalised practice, including tailored explanations and follow-up prompts that scaffold deeper understanding. Together, these features highlight RiPPL’s integration of generative AI to strengthen learning, feedback, and learner agency at scale.

companions encourage learners to justify design choices, link concepts to prior knowledge, and refine explanations (e.g., “How does this connect to the foundational idea introduced earlier?”) [86]. During peer evaluation, companions scaffold high-quality feedback by modelling constructive comments, identifying missing rationale, and encouraging reviewers to make comments more actionable. These supports deepen cognitive engagement without replacing the epistemic work required for learning.

Spaced retrieval and practice through personalised cycles of engagement. RiPPL embeds the principles of spaced retrieval and spaced practice by recommending activities based on learners’ mastery trajectories. Retrieval practice—particularly when effortful—is one of the most powerful mechanisms for strengthening long-term memory [93]. RiPPL ensures that opportunities for retrieval are distributed across time by interleaving practice with cycles of content creation and peer review, enabling repeated engagement with concepts from different angles. AI companions further reinforce these mechanisms by providing adaptive, personalised support during practice. They may prompt learners to recall related principles before offering hints, encourage justification of chosen answers, or surface past misconceptions to stimulate reflection. Companions can also generate follow-up challenges, recommend revisiting

earlier created or reviewed resources, or propose consolidation tasks when recurrent patterns of error are detected. These behaviours ensure that retrieval remains active, distributed, and conceptually relevant, supporting deeper consolidation and transfer.

6.2.3 Adaptive Foundations of RiPPLE

Central to RiPPLE’s design is the integration of learner models that drive personalised recommendations and shape AI companions’ interactions. The platform captures rich digital traces—including performance, contribution history, peer evaluations, time-on-task, and patterns of misconception—to form multi-dimensional representations of each learner’s knowledge, progress. These learner models power RiPPLE’s adaptive recommendation engine, which selects suitable practice items, identifies gaps in conceptual understanding, and schedules review activities at optimal intervals [55]. When integrated with generative AI companions, these models enable the system to deliver personalised, contextualised support. Beyond cognitive modelling, RiPPLE captures contextual features of the learning experience such as the difficulty of contributed questions, peer ratings, or the diversity of resources encountered. Learning companions use these contextual signals to determine the granularity of explanations, when to prompt reflection, or when to escalate support. This harmonises micro-level adaptivity with macro-level guidance, enabling companions to learn from and evolve with each student’s trajectory.

6.2.4 Responsible Design Foundations in RiPPLE

RiPPLE’s implementation of learning companions is grounded in key elements of the responsible design principles outlined earlier, with particular emphasis on transparency, explainability, and meaningful human oversight. These foundations ensure that AI support enhances learning without undermining learner agency, instructional control, or pedagogical integrity.

Explainable learner modelling. Aligned with the principles of XAI-ED [53], RiPPLE enables learners to understand why the system makes particular recommendations or highlights specific areas for review. Learning companions provide clear, context-sensitive explanations of their decisions, such as: “I suggested revisiting this topic because your last three attempts showed the same misconception.” These explanations make the system’s reasoning visible and intelligible, helping learners calibrate their trust in the platform and encouraging active engagement in metacognitive monitoring. By situating feedback within the learner’s behavioural and conceptual history, explainability becomes a pedagogically meaningful scaffold rather than a purely technical disclosure.

Instructor oversight and auditability. Human oversight is preserved through comprehensive instructor dashboards that surface all learner-generated content, AI-supported feedback exchanges, and moderation outcomes. Educators can review resources that have been flagged for potential issues, and identify recurring misconceptions across the cohort. This auditability allows instructors to intervene when needed, refine instructional strategies, and maintain authority over judgements that carry pedagogical or assessment stakes. Importantly, the design ensures that automated systems do not make unilateral high-stakes decisions; instead, AI acts as a support layer that augments, rather than replaces, human expertise.

6.3 CodeHelp: An AI-Powered Tutor for Scaffolded Programming Support

6.3.1 Overview of CodeHelp

CodeHelp is an LLM-powered web application designed to provide scalable, on-demand support to programming students without revealing code solutions [65]. This design is supported by recent research showing that students in programming courses value fast access to help but prefer scaffolded guidance over direct answers [33]. CodeHelp is one of the flagship applications built on the Gen-Ed (Generative AI for Education) framework, an open-source infrastructure for constructing educational web applications. Gen-Ed provides authentication (including single sign-on), LMS integration via LTI, class enrollment and role management, administrative dashboards, and data export capabilities.

CodeHelp provides two primary forms of support. The first is the ‘Code Query’ feature, which allows students to request help with programming issues such as debugging, interpreting error messages, or understanding conceptual aspects of their code. Students submit structured inputs that include the programming language, relevant code, error messages, and a description of their question. Responses are generated through a multi-stage prompting pipeline designed to avoid providing solution code. This pipeline includes (1) a sufficiency check to determine whether the student’s query contains adequate information, (2) generation and scoring of multiple candidate responses, and (3) a post-processing step that removes any code blocks from the response. This design ensures that students receive conceptual guidance without direct answers. An example of a guardrailed response from CodeHelp is illustrated in Figure 5(a).

Building AI Companions that Prioritise Learning over Performance

Response *Remember: It will not always be correct!*

Hello! It seems like you're trying to calculate the sum of all values surrounding a central value in a 10x10 two-dimensional array. The issue you're encountering is likely due to the fact that your function does not account for the boundaries of the array.

When `row` or `col` is near the edges of the array, your loop variables `i` and `j` can go out of bounds, which means they can become less than 0 or greater than 9. When you try to access `values[i][j]` with `i` or `j` out of bounds, you're accessing memory that is not part of the array, which can lead to unpredictable results or even crashes.

To fix this, you need to ensure that `i` and `j` stay within the bounds of the array. You can do this by adding checks within your nested for loops to continue to the next iteration if `i` or `j` is less than 0 or greater than 9. This way, you'll only sum up the values that are within the array.

Remember, in C, arrays are zero-indexed, which means the first element is at index 0 and the last element is at an index one less than the size of the array. So for a 10x10 array, valid indices are from 0 to 9.

I hope this helps! Let me know if you have any other questions.

Was this helpful? Yes No

(a) Example response from CodeHelp, providing next step guidance without revealing a direct code solution

Learning Objectives

Regenerate draft learning objectives.

This will overwrite existing objectives and questions.

1. Declare and initialize pointer variables using the address operator (&) in C, and understand their role in storing memory addresses. Only focus on the int type (so int *)
2. Dereference initialized pointer variables using the dereference operator (*) to access or modify the value pointed to, while avoiding dereferencing null or uninitialized pointers. Only focus on the int type (so int *)
3. Explain the difference between pass-by-value and pass-by-reference parameter passing in C, and how pointers enable pass-by-reference.
4. Declare and initialize two-dimensional arrays, and use nested loops to systematically access and manipulate their elements.
5. Pass arrays and pointers as parameters to functions, understanding how this enables functions to modify original data and the requirements for passing multi-dimensional arrays.

Topic: Loops, Conditionals, Arrays, Functions and Parameter passing

Tutor Hi! We're going to explore some fundamental programming concepts in C: loops, conditionals, arrays, functions, and how parameters are passed. To start, could you tell me what you already know about using conditional statements like `if` and `else` in C? For example, have you worked with boolean expressions or conditions before?

You

Tutor's plan & your progress	
IN PROGRESS	Evaluate boolean expressions in C and use conditional (<code>if</code> / <code>else</code>) statements to execute code blocks.
NOT STARTED	Implement while and for loops to perform repetitive tasks.
NOT STARTED	Declare, initialize, and access elements of one-dimensional arrays.
NOT STARTED	Use loops to iterate over the elements of the a one-dimensional array, performing operations like summation and searching
NOT STARTED	Define and call functions with parameters and return values.

(b) Learning outcomes inferred from course material

(c) Highlighting learner progress with a 'guided tutor'

Figure 5: Illustrative examples from the CodeHelp platform. Panel (a) illustrates a guardrailed response in the 'Code query' mode after a student submitted code containing a bug. Panel (b) illustrates the instructor configuring a 'guided tutor' session. After uploading relevant course material, a set of learning outcomes are generated for review. These can be edited directly or regenerated. Each learning outcome is accompanied by a step-by-step associated plan (not shown in the figure), which is also generated by CodeHelp for the instructor's review, and which scaffolds the learner's interaction with the tutor. Panel (c) shows the start of a 'guided tutor' session, with a table that highlights the learner's progress through the plan.

The second feature is the 'Tutors' system. CodeHelp supports both open-ended 'Inquiry Chats', where students initiate discussions about topics of interest, and 'Focused Tutors', which are configured by instructors around specific learning objectives. For 'Focused Tutors', instructors upload course materials and CodeHelp automatically proposes learning objectives and associated assessment questions. Instructors edit and refine these suggestions prior to deployment. During an interactive session with a student, the tutor guides the student through the objectives interactively, probing for understanding and requiring evidence of mastery before progressing. Figures 5(b) and (c) illustrate examples of the learning objectives generated from provided course material, and the beginning of a 'guided tutor' session, respectively. Responses are deliberately framed in a positive and encouraging tone, modelling supportive human tutoring practices [66]. CodeHelp shares similarities with other programming assistants such as CodeAid [52], particularly in its use of guardrails to prevent direct code generation.

6.3.2 Pedagogical Foundations of CodeHelp

Deep and Interactive Learning CodeHelp is intended to function similarly to a human tutor by engaging in guided, back-and-forth interactions with students rather than simply delivering answers. This approach aligns with the ‘Deep and Interactive Learning’ principle by promoting active cognitive engagement rather than passive consumption of solutions. With the ‘Focused Tutors’ feature, students engage in dialogue as they step through questions posed by the tool which relate to the learning outcomes approved by the instructor. As they do so, the system probes for understanding, asks follow-up questions, and adapts its explanations to the student’s responses. This feature requires students to demonstrate understanding before they progress. In the ‘Code Query’ feature, students are free to ask their own questions and they receive conceptual explanations and guidance rather than complete implementations. This requires students to interpret the guidance provided, rather than passively consume solutions.

Guided Scaffolding and Productive Struggle A defining feature of CodeHelp is its explicit avoidance of solution code when students are seeking help in the ‘Code Query’ mode. The multi-stage prompting pipeline that CodeHelp uses serves as a pedagogical guardrail to support understanding rather than bypassing it. The use of such guardrails also introduces an element of productive challenge, where students have to interpret the feedback provided and apply it within their own programming context. Responses are deliberately phrased in a positive and encouraging tone, in order to sustain motivation.

Learning to Learn and Metacognitive Development CodeHelp provides a query interface that is divided into four areas: programming language, relevant code fragment, error message, and free form question. This structure is designed to “provide guidance to students about what information is typically needed for an effective query” and to give “students an opportunity to practice asking technical questions, providing the necessary relevant context” [65]. More importantly, CodeHelp includes a ‘sufficiency check’ as part of the prompting pipeline, which determines if all essential information for answering the query has been provided by the student. If any important information is missing, rather than providing a potentially inaccurate response, CodeHelp will respond to the student’s question by asking for the missing information. This combination of the structured input field and the sufficiency check helps students form precise questions and thus models an important aspect of effective help seeking, which is an important metacognitive skill.

The ‘Focused Tutors’ mode further supports metacognitive development by prompting students to reflect on their understanding as they work through instructor-defined learning outcomes. Students are asked open-ended questions that probe conceptual understanding, and they are expected to articulate their reasoning before progressing through to the next step in the guided tutor plan. When misunderstandings arise, the tutor provides corrective feedback and simplified probing rather than simply supplying answers. This structure encourages learners to monitor their own understanding and identify gaps in their knowledge, and thus helps to support the development of essential self-regulation skills. This mode can be incorporated into a course by instructors for regular use, for example to facilitate end-of-week review sessions.

Contextual and Authentic Learning By design, the ‘Code Query’ feature of CodeHelp supports student learning in the context of authentic programming tasks. Students use the tool while working on real assignments, often outside class hours. Usage data from a semester-long deployment in an introductory computer and data science course (n = 52) showed sustained engagement over 12 weeks, with more than 2,500 student queries submitted [96]. Most queries focused on debugging and implementation questions that were directly tied to the course assignments, demonstrating highly contextual use. The study also found a modest positive correlation between overall tool usage and course performance, suggesting that engagement with the tool did not undermine learning and may have supported it.

Students typically use the ‘Code Query’ feature of CodeHelp after encountering a difficulty or error when programming, such as an error message or unexpected behaviour of their program. This prompts them to reflect on what might be going wrong. CodeHelp responds with conceptual explanations and high-level guidance, after which students return to their code to revise and test their approach. This iterative pattern of attempt, reflection, refinement, and re-experimentation mirrors Kolb’s experiential learning cycle (see Section 3.4) and situates learning within active problem solving.

6.3.3 Adaptive Foundations of CodeHelp

Although CodeHelp demonstrates strong alignment with the pedagogical foundations detailed in Section 3, its adaptive capabilities are currently limited. Nevertheless, it illustrates how pedagogically grounded systems can provide useful learning support even before full adaptivity is implemented.

As outlined in Section 4.1, the ‘Capture’ stage of the adaptive cycle describes the importance of systematically recording digital traces in order to enable learner modelling and provide adaptive support. The CodeHelp system does capture detailed interaction data, including all queries and responses, which can be reviewed and exported by instructors. These conversational exchanges between the student and the AI model form digital footprints of learners’ problem-solving processes. Analyses reveal substantial variation across students in the kinds of help they seek, and their patterns of

engagement [96]. This variation is particularly valuable for adaptive systems as it provides a richer signal for modelling. Translating these captured traces into learner models that inform more dynamic adaptive support across sessions remains an important direction for future work.

The adaptation that does occur currently in CodeHelp is limited to within individual sessions. For example, in the ‘Focused Tutors’ mode, the system adjusts its responses and the probing questions that it asks based on how students answer. However, this adaptation does not persist across sessions, and the system does not automatically refine its behaviour based on accumulated data. Instructors may manually adjust learning outcomes or prompts in response to observed patterns, but this currently remains human-mediated rather than automatic.

6.3.4 Responsible Design Foundations of CodeHelp

Security and Privacy The architectural design of CodeHelp provides support for important aspects of the Responsible Design pillar of the framework. As an open-source system that can be self-hosted, institutions have the option to retain control over data storage and processing. The platform integrates with LMS environments via LTI, aligning with established institutional security practices. Importantly, even the hosted version of CodeHelp allows instructors to configure the AI backend using any OpenAI-compatible REST endpoint. This design means that instructors and institutions have the option to use their own managed or locally hosted AI models. However, responsibility for secure configuration ultimately rests with deploying institutions.

Transparency and Inclusion CodeHelp promotes transparency through the public availability of its source code and prompting pipeline, which means that the guardrail mechanisms can be reviewed. Student-facing responses include a notice that acknowledges outputs may not always be correct, highlighting the limitations of LLM-based systems. One of the main motivations for the CodeHelp project was to lower barriers to help-seeking by offering private, always-available assistance, which can support inclusion given that in-person support often fails to reach all students equally [98]. Research suggests that some learners do feel more comfortable interacting with the system than approaching instructors or TAs directly [65].

Accountability Instructor oversight is a core part of the system design. Dashboards allow instructors to monitor usage patterns and review and export queries and responses. With the ‘Focused Tutors’ mode, instructors retain control over the suggested learning outcomes that CodeHelp generates from provided course material, and they can edit them to align with their pedagogical goals (see Figure 5(b)). To further customise model responses, instructors can define an ‘avoid set’ which is a set of keywords that should be avoided in any AI response to the student. In addition, after each response the system asks students to indicate whether the response was helpful (see Figure 5(a)), creating an auditable feedback signal that can support review and ongoing refinement. While the actual interactions are generated dynamically by the underlying language model, these mechanisms help to support instructor oversight of their learners’ interactions.

6.4 JeepyTA: An AI-Powered Teaching Assistant for Course Discussion and Formative Feedback

6.4.1 Overview of the JeepyTA

JeepyTA is an AI-driven teaching assistant developed by the Penn Center for Learning Analytics at the University of Pennsylvania, first launched in Fall 2023. The system employs a multi-turn conversational architecture powered by large language models and is model-agnostic, capable of operating with various LLMs including GPT, Llama, and DeepSeek. Deployments have utilized successive OpenAI GPT models, from GPT-3.5 Turbo through GPT-5.

As of late 2025, JeepyTA has been deployed across nearly 20 courses at four institutions in the United States and Singapore. Integrated into the open-source Flarum discussion forum where it appears as a distinctly marked AI Teaching Assistant, it can also be integrated into discussion forums in other platforms such as Canvas. A Progressive Web App architecture provides mobile accessibility with email and push notifications.

JeepyTA’s capabilities span multiple pedagogical functions: answering logistical questions, participating in discussions, delivering targeted feedback on essays, projects, and code, and supporting brainstorming. To ensure alignment with specific course contexts, instructors prime JeepyTA with chosen reference materials: syllabi, textbooks, readings, and past feedback examples. These resources are embedded through a retrieval-augmented generation (RAG) workflow. Instructors can customize behavior by task, determining which areas JeepyTA addresses and the required oversight level, including options for human-in-the-loop review before responses become visible to students.

Figure 6 depicts JeepyTA’s core workflow. The instructor selects a set of pedagogical functions for a specific course. When the student interacts with the relevant activities – posting a question to the discussion forum or submitting an assignment for feedback – JeepyTA generates a response. In some cases, the instructor reviews that feedback; in other cases the student receives it immediately. Next, the student chooses how to act based on the response, perhaps asking a

follow-up question, submitting a revised assignment, or ending the discussion thread. This sequence of events happens multiple times per student, per course.

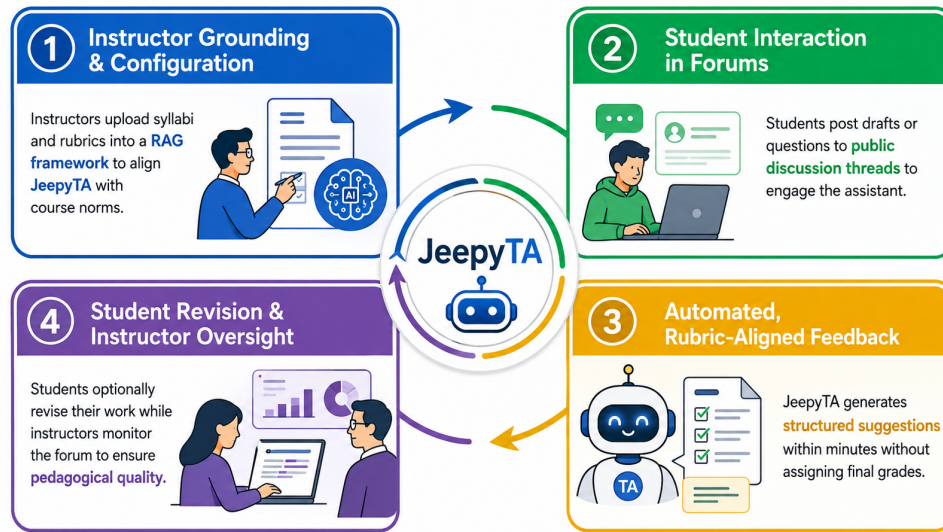


Figure 6: Overview of the JeepyTA workflow

6.4.2 Pedagogical Foundations of JeepyTA

Deep and Interactive Learning. JeepyTA engages students through sustained dialogue. When responding to reflections on readings and lectures, it acknowledges contributions, reinforces key ideas, and connects insights to course themes. It clarifies concepts by summarizing arguments and citing specific readings. When students introduce interesting perspectives, JeepyTA poses follow-up questions to encourage further discussion, supporting the iterative meaning-making process central to deep learning.

In brainstorming mode, JeepyTA serves as a collaborative partner. In a Games and Learning course, students consulted it to propose educational uses for video games through structured play journal assignments, with JeepyTA scaffolding connections between gameplay experience and pedagogical application.

Guided Scaffolding. When delivering feedback on student essays and projects, JeepyTA evaluates both higher-level conceptual elements—such as whether students have employed course theories in their arguments, detailed solution limitations, and appropriately addressed stakeholder needs—and lower-level aspects including argument clarity, evidence use, structure, and writing quality. In its prompt, JeepyTA is asked to “provide actionable insights rather than shallow suggestions,” offering concrete guidance.

For programming, JeepyTA identifies errors and suggests corrections without providing complete solutions, helping students develop diagnostic skills and problem-solving capacity. When students describe issues vaguely, JeepyTA asks clarifying questions about error messages, intended functionality, or steps already attempted, scaffolding the troubleshooting process itself.

The system’s scaffolding approach incorporates instructor oversight mechanisms that balance immediacy with accuracy. Instructors can configure whether JeepyTA’s responses appear immediately or await human review, adjustable by discussion category. This flexibility allows graduated release of autonomy. In practice, instructors often apply tighter control for conceptual questions, while enabling immediate formative feedback for assignments to support rapid iteration.

Learning to Learn and Higher Order Learning. JeepyTA supports metacognitive development through feedback and reflective prompts. When providing essay feedback, the system comments on strategic aspects of writing such as how arguments are constructed and how evidence is marshalled, making these higher-order processes visible and discussable. In debugging support, JeepyTA suggests higher-order strategies such as adding diagnostic print statements, checking variable values systematically, or decomposing complex functions into testable components.

However, the system remains primarily responsive rather than proactive in cultivating metacognitive awareness. While JeepyTA explains reasoning after student attempts, it rarely prompts students to articulate their own learning goals,

monitor progress, or reflect on the evolution of their understanding. The weekly discussion summaries it offers within some courses (again at instructor discretion) represent one mechanism for prompting reflection, by highlighting key themes, recurring arguments, and unresolved questions. However, these summaries serve instructors and the collective more than individual metacognitive development.

Contextual Learning. JeepyTA situates learning authentically through course-specific grounding. In a course on Educational Data Mining, for example, JeepyTA was explicitly instructed to prioritize methodologically appropriate techniques for the domain rather than generic approaches that might be acceptable elsewhere.

Another mechanism used by JeepyTA for contextual learning is its persona functionality. In a course on cultural foundations for teaching and learning, JeepyTA adopted personas based on lived experiences of individuals from historically underrepresented groups, defined through description (name, role, context, cultural background), situation (interaction role, task, participants), and instruction (tone, detail, specific elements). By engaging with these personas—such as “Felipe,” a Mexican American teacher educator, or “Claire,” a Hmong American educator—preservice teachers encountered concrete examples of culturally relevant pedagogy, including specific strategies for incorporating funds of knowledge into classroom settings.

6.4.3 Adaptive Foundations of JeepyTA

JeepyTA’s current implementation represents an early stage in realizing Section 4’s adaptive foundations, with stronger development in some dimensions than others.

Capture. The system logs comprehensive conversational data: student queries, JeepyTA responses, instructor modifications, and follow-ups. The RAG architecture tracks which materials are retrieved per query, providing implicit evidence of which concepts students are engaging with or struggling to understand. Additionally, when JeepyTA provides essay feedback or debugging support, the student work itself can reveal conceptual understanding, strategic approaches, and error patterns. Recent analyses have leveraged these data to study how students adapt work following feedback.

However, capture remains limited compared to Section 4.1’s vision. The system does not currently integrate data from beyond the forum, such as performance or time-on-task in other course components, or multimodal data.

Model. Learner modeling capability is minimal. JeepyTA maintains no persistent representations of individual students’ knowledge, misconceptions, strategies, or affect. Each discussion thread is treated independently and is not supplemented by a cumulative student model.

This represents a significant gap relative to this paper’s adaptive foundation vision. The LLM’s conversational memory within a specific discussion thread provides some continuity, but no other measurement of knowledge, affect, metacognition, or other key quantities is currently present.

Adapt. JeepyTA’s adaptive behavior occurs at the level of RAG content retrieval and situational prompt design rather than deep personalization. The system can be configured with different behavioral rules by discussion category, but this is an instructor-level design choice rather than true system adaptivity.

Evolve. JeepyTA incorporates mechanisms for system-level improvement, but these are human-mediated rather than automated. Instructors review, approve, edit, or discard responses, with the option to remember these corrections for future interactions. This creates feedback loops where the knowledge base and behaviors improve based on instructor input. When instructors note recurring issues, such as JeepyTA defaulting to inappropriate problem-solving approaches or missing course-specific conventions, prompts can be refined to address these patterns. For example, instructors have iterated prompts for feedback and brainstorming based on prior semesters’ experiences, but this remains in the realm of learning engineering rather than adaptive learning.

What remains absent is the kind of automated, continuous evolution described in Section 4.4, where the system employs A/B testing, multi-armed bandit algorithms, or reinforcement learning to discover which pedagogical strategies work best for which students in which circumstances.

6.4.4 Responsible Design Foundations in JeepyTA

JeepyTA’s implementation demonstrates attention to several dimensions of responsible AI design, with particular emphasis on transparency and human oversight.

Transparency. JeepyTA implements several transparency mechanisms that align with the principles outlined in Section 5.2. Most fundamentally, the system is distinctly marked as an AI Teaching Assistant within the forum, ensuring students understand they are interacting with an automated system rather than a human TA or instructor. The system also provides transparent explanations in many interactions. JeepyTA often cites course materials explicitly. This

allows students to verify information against original sources and understand the evidential basis for responses. When providing essay feedback, JeepyTA grounds its comments in rubric criteria, clarifying which standards are being applied and why particular aspects receive attention.

However, transparency is incomplete. It does not maintain open learner models that would make its interpretations visible for inspection and negotiation. Like many LLM-based systems, the underlying architecture remains a “black box”—neither students nor instructors can trace how inputs and retrieved materials transform into outputs.

Accountability. JeepyTA is designed with a commitment to human oversight. The configurable review mechanism—allowing instructors to determine whether responses require approval before posting, adjustable by category—preserves instructor authority while enabling rapid feedback where appropriate. JeepyTA handles high-volume, routine queries and provides swift formative guidance; human educators remain responsible for conceptual instruction and high-stakes decisions. The system preserves audit trails of all interactions, supporting retrospective analysis, continuous improvement, and accountability when issues arise.

Inclusivity. JeepyTA incorporates several inclusivity-oriented features, though this remains an area for further development. Its 24/7 availability supports equitable access for students with non-traditional schedules, caregiving responsibilities, or different time zones. Students may also receive faster responses from human instructors and TAs, who can focus attention on situations requiring human expertise (Liu et al., in press). JeepyTA additionally interacts in multiple languages, leveraging the multilingual capabilities of its underlying LLMs.

Beyond this, the persona functionality explicitly centers diverse cultural perspectives and funds of knowledge. By grounding personas in lived experiences of individuals from historically underrepresented groups, the system exposes learners to a wider range of worldviews. This approach is based on the premise that inclusive AI should not only avoid bias but also actively support engagement with diverse ways of knowing and being.

6.5 Recast: A university platform for AI Companions

6.5.1 Overview of Recast

The University of Technology Sydney (UTS) is a public institution serving approximately 50,000 students studying in hybrid and fully online modes, in Sydney and abroad. UTS has developed a platform called *Recast* to support the design and publishing of AI Companions, (termed Recast Assistants) for staff and students. Recast provides an Assistant design interface enabling staff to design and test system prompts, adjust the usual dialogue parameters such as temperature and strictness, and optionally ground the Assistant in a folder of documents for Retrieval Augmented Generation (RAG) if it is desired that its answers are sourced only from that curated corpus (hence, Assistants do not perform dynamic internet searches: Microsoft Copilot is provided for this). Role-based permissions manage student and staff access to Assistants via their university account, making this a formal part of the university’s learning technology ecosystem that can be used for formal assessments. Assistants can be set to offer voice input/output, while custom prompting reflects the specific pedagogy, topics and learning outcomes of their course, or a specific assignment for which the Assistant was designed.

6.5.2 Pedagogical foundations of Recast

Since Recast is a platform for designing and deploying diverse types of Assistant, there is no single tool to describe. Rather, families of Assistant are emerging, with correspondingly diverse pedagogical bases, conducting distinct kinds of dialogue. Some of these are introduced next, mapped to two of the processes in the pedagogical foundation introduced above. As can be seen, these are focused less on coaching mastery of curriculum concepts, and more on higher order, contextualised skills.

Learning to learn & higher order learning. *Reflection Assistants* conduct a strongly scaffolded student reflection on some aspect of their course. An example would be reflection on the progress students have made on goals they set for themselves at the start of semester, prior to submitting a written reflection. These submitted reflections can then be analysed using an LLM for evidence of students’ affective engagement and belonging [90]. *Critical Thinking Assistants* use a form of Socratic questioning about the implicit premises behind students’ questions, to help students check their assumptions and refine their research questions [63]. This has been used in courses including Data Science and Sustainable Futures, as well as by PhD students and academics.

Contextual learning. *Role-Play Assistants* provide authentic, experiential learning through voice-based role-play simulations [63]. These enable students to practice conducting challenging conversations within the dynamics of a verbal interaction (in contrast to slower, written dialogue). At the time of writing, we have such Assistants integrated

into Health and Law assessments, whose roles and conversations are, as one would expect, very different. Snapshots of these and other Assistants are provided online,¹ and empirical studies are now underway.

6.5.3 Adaptive foundations of Recast

Recast Assistants are not adaptive in the sense that an ITS continually updates a persistent learner model with the student's level of competence mapped to a detailed curriculum model of requisite knowledge and skills. However, within the capability of an LLM with custom prompting and grounding, the Assistants *capture* all interactions persistently (opening the possibility of running advanced analytics on the transcripts), and *adapt* in the sense that their responses are contingent on what a student says, guided and constrained by diverse forms of pedagogically-tuned system prompt.

6.5.4 Responsible design foundations of Recast

In conceiving, designing, implementing and evaluating Recast as a trustworthy institutional platform for AI Companions, it has become evident that responsible design is a complex, sociotechnical process, interweaving technical, user experience, policy and governance practices. The following sections explain the choices UTS has made plus rationale. However, this remains a turbulent landscape: as UTS gains governance experience in deploying Recast at scale, as empirical evidence builds on student and educator experiences, and as technological capabilities and limitations become apparent, these design decisions will be under regular review.

Security and privacy. Recast had to pass rigorous cybersecurity and architecture audits to be part of the UTS enterprise computing infrastructure. LLMs are hosted securely within an enterprise AI vendor with all data storage within state, while user authentication is handled by the institutional identity and permissions layers, drawing from the enrolment database to control which Assistants a student sees. All Assistant transcripts are archived, since UTS considers itself accountable for how its platforms are being used, must be able to verify performance for quality assurance, and investigate any incidents. In institutional ethics-approved research projects, student identifiers are not included in transcripts and replaced with new identifiers to enable dataset merging.

Transparency on transcript privacy. A particular context for transparency concerns the privacy of Recast transcripts. Unlike Khanmigo, our students are adults, and the choice was made that the teaching team cannot by default see student transcripts. Students are informed that all Recast sessions are logged, but that their individually identifiable chats are not being monitored by academics/tutors. The rationale for this is it would likely disincentivise student adoption, or impair engagement, given the evidence that students value chatbots as private support, avoiding fear of embarrassment or trepidation in approaching staff [46]. A teacher dashboard provides cohort-level quantitative indicators of engagement, which in the future could include more advanced, research-informed transcript analytics tuned to the nature of the dialogue. There are two exceptions: (i) the lead academic can view deidentified transcripts if students have given informed consent to participate in an ethics-approved research project; (ii) tutors can view identifiable transcripts when the Recast session is a formal requirement for an assignment in which the transcript is a primary artifact for assessment (e.g., a role-play dialogue).

Accountability. Firstly, there is accountability for AI ethics. In 2021, we initiated a set of student and staff “EdTech Ethics” consultations, which proved to be a rewarding demonstration of the value that a deliberative democracy process can bring to organisation-wide deliberation [101]. Subsequently, UTS formed a cross-university AI Operations Board, the governance body that implements our AI Operations Policy, which in turn builds on the state's AI ethics framework. Proposals for Recast Assistants are governed by the same review process that covers all new AI applications within UTS. This introduces a second form of accountability, that we have confidence in an Assistant's trustworthiness. Recast quality assurance requires developmental testing by the teaching team, plus automated testing that exhaustively verifies appropriate responses to a range of likely prompts, wellbeing alerts, as well as jailbreaking efforts. Finally, there is accountability for student wellbeing. Content filtering is designed to identify concerning material from either the student or LLM, at which point Recast declines to engage in further discussion. If student distress is detected, guardrails in the system prompt link students to UTS Counselling. Filters can be adjusted at the individual Assistant level, so that students can legitimately discuss sensitive topics for their studies (e.g., in Forensics, or Social Work).

Inclusivity. It was a requirement that the user interface meet W3C accessibility standards, which underwent testing by the UTS accessibility specialist team. However, there are degrees of accessibility, and there remain aspects that should be improved.

To summarise, the Recast case demonstrates that in order to implement the responsible design framework principles in an AI Companion platform operating in the enterprise IT at institutional scale, technical choices under the control of developers alone are critical, but insufficient: they must also align with transparent communication by the teaching

¹UTS:CIC GenAI Apps & Analytics: <https://cic.uts.edu.au/tools/genai-apps-and-analytics>

Building AI Companions that Prioritise Learning over Performance

Case study	Pedagogical foundation <i>How students learn with AI</i>	Adaptive foundation <i>How AI learns about students</i>	Responsible design foundation <i>Integrity & human values</i>
Khanmigo <i>An AI-Powered Tutor for Global Skill Practice</i>	[Strong] Moved away from pure Socratic questioning after transcripts revealed student frustration; redesigned around hints and worked examples. ICAP framework measures cognitive engagement at scale.	[Partial] Heuristic mastery tracking chosen over probabilistic modelling for transparency. Skill-level mastery injected into prompts. Interest-based personalisation trialled but yielded no gains.	[Strong] Full teacher and parent transcript visibility; AI moderation for harmful content; contractual data-deletion with model providers. K-12 learner context drives design choices.
RiPPLE <i>An AI-Powered Co-Creation and Peer Learning Platform</i>	[Strong] Elaboration through student content creation; spaced retrieval across creation–evaluation–practice cycles. AI companions scaffold peer feedback without replacing epistemic effort.	[Strong] Most fully realised adaptive cycle: multi-dimensional learner model integrating performance, misconceptions, and peer ratings. Adaptive engine governs both item sequencing and companion micro-decisions.	[Strong] XAI-ED explainability: companions cite specific evidence behind each recommendation. Instructor dashboards surface all AI exchanges. High-stakes decisions remain human-gated.
CodeHelp <i>An AI-Powered Tutor for Scaffolded Programming Support</i>	[Strong] Multi-stage pipeline withholds solution code by design. Sufficiency check teaches effective help-seeking as a metacognitive skill. Focused Tutors require demonstrated understanding before progression.	[Partial] Within-session adaptation only: probing questions adjust to student responses but no cross-session learner model is maintained. Translating captured logs into persistent adaptation is the key next step.	[Partial] Open-source codebase and self-hosting give institutions data control. Instructor dashboards and student helpfulness ratings provide accountability. Governance relies on institutional configuration.
JeepyTA <i>An AI-Powered Teaching Assistant for Discussion and Formative Feedback</i>	[Strong] Broad range: dialogue on readings, rubric-grounded essay feedback, programming support without full solutions, and culturally situated persona-based learning. Contextual and higher-order learning are particularly well developed.	[Emerging] No persistent learner model. Adaptation limited to RAG retrieval and situational prompting. Evolution is human-mediated through instructor prompt refinement rather than automated.	[Partial] Clearly marked as AI; responses cite course materials. Configurable human-in-the-loop review by category preserves instructor authority. Audit trails maintained; architectural explainability absent.
Recast (UTS) <i>AI-Powered Higher-Order and Contextual Learning</i>	[Strong] Distinct assistant families target higher-order and contextual learning: Reflection Assistants, Critical Thinking Assistants using Socratic questioning, and voice-based Role-Play Assistants for Health and Law.	[Emerging] No persistent learner model; adaptation is LLM-contingent within sessions. All transcripts archived as a deliberate foundation for future analytics pipelines.	[Strong] Most institutionally elaborate: cybersecurity audits, AI Operations Board, automated jailbreak testing, W3C accessibility, and student ethics consultations. Transcript privacy by default for adult learners—the deliberate inverse of Khanmigo.

Figure 7: Synthesis of case studies mapped against the three framework pillars.

team to students about who can see what, and with the institution’s policy and governance mechanisms. There is no context-independent blueprint for how to implement AI companions responsibly, but this case demonstrates how and why one institution made its choices.

Pedagogical, Adaptive, and Responsible Foundations in Practice

Three patterns emerge from the case studies, aligned with the framework (see Table 7). First, the pedagogical foundation is the most mature across all cases, but different key mechanisms are used by the tools. Khanmigo and CodeHelp focus on avoiding the generation of direct solutions in order to preserve productive struggle; RiPPLE shifts learners into generative roles as authors, peer reviewers, and active practitioners; JeepyTA embeds support across a range of activities that are highly course-specific and contextualised; and Recast extends the idea of a learning companion beyond content

and into reflection, critical thinking, and authentic role-play. Across these cases, a central pedagogical design decision is around which cognitive work should remain with the learner.

Second, the adaptive foundation reveals the greatest variation in maturity. RiPPLE most closely approximates the full adaptive cycle, with persistent learner modelling and personalised recommendation mechanisms already embedded in the platform. Khanmigo uses a narrower form of adaptivity by incorporating mastery information about the current and prerequisite skills into tutoring interactions. CodeHelp, JeepyTA, and Recast all capture some traces of learner interaction and adapt within sessions or contexts, but the way they adapt remains largely human-mediated rather than automatic via learner models. Persistent, longitudinal adaptivity remains the most significant gap between current implementations and the full vision of an AI learning companion.

Third, the responsible design foundation is highly context-dependent. Khanmigo’s full conversation visibility is appropriate to a school-based context, whereas Recast’s transcript privacy by default reflects a university context in which adult learners may need private space to experiment, and ask questions that may reveal a lack of understanding or uncertainty. CodeHelp and JeepyTA provide for instructor configurability and auditability, while RiPPLE emphasises explainable learner modelling and instructor oversight of learner-generated content. These differences show that responsibility depends very much on the learner and institutional contexts in which the companion operates.

The five case studies demonstrate that current AI-powered educational tools are beginning to progress beyond generic task-completion tools, although progress with respect to the foundations of our AI companion framework is uneven. Their strongest common contribution lies in pedagogical design where each tool attempts, in different ways, to keep learners cognitively active and reflective. One challenge ahead is to close the adaptive loop, and using persistent evidence about learners to personalise support without compromising privacy, agency, or human oversight. Ultimately, learning-oriented AI should be judged by how deliberately it supports the processes through which durable learning develops.

7 Conclusion

This paper has argued that the growing integration of LLMs into education demands more than better prompting or pedagogical guardrails applied to tools built for other purposes. It requires a fundamental reconceptualisation of what AI in education is for. The learning-performance paradox reveals that AI optimised for task completion can actively undermine the processes through which durable learning occurs. The response we have proposed is a distinct class of AI agents, AI learning companions, designed to be embedded in educational environments with an explicit mission to prioritise learning over performance. The framework we have presented organises companion design around three interrelated foundations: a pedagogical foundation concerned with how students learn with AI, an adaptive foundation concerned with how AI learns about students, and a responsible design foundation concerned with how companions act with integrity and uphold human values.

The five case studies, each authored by researchers who led or made significant contributions to the tools described, provide firsthand insight into how these foundations play out in practice across diverse educational contexts, levels, and design approaches. They collectively demonstrate that deliberate pedagogical design can meaningfully distinguish learning-oriented AI from generic task completion tools, and that responsible design looks different depending on whether learners are minors or adults, whether tools are purpose-built or platform-based, and whether the goal is content mastery or higher-order development. Persistent adaptivity remains largely emerging across most cases, pointing to the most important directions for future work: how persistent learner modelling can be developed without compromising privacy or learner agency; how delayed retention and transfer can be built into evaluation frameworks as standard practice rather than an afterthought; and how governance models can be adapted and scaled across diverse institutional and regulatory contexts.

Ultimately, the promise of AI learning companions lies not in replacing human teaching or accelerating task performance, but in cultivating learners who are more reflective, more metacognitively aware, and better equipped to learn independently in an AI-rich world. The shift from optimising for performance to cultivating learning is not a minor adjustment to how we deploy AI in education. It is a fundamental reorientation of what we ask AI to do for learners, and for education itself.

References

- [1] G. Abdelrahman, Q. Wang, and B. Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

- [2] S. Abdi, H. Khosravi, S. Sadiq, and D. Gasevic. Complementing educational recommender systems with open learner models. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 360–365, 2020.
- [3] V. Aleven, B. McLaren, I. Roll, and K. Koedinger. Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In *International conference on intelligent tutoring systems*, pages 227–239. Springer, 2004.
- [4] V. Aleven, E. A. McLaughlin, R. A. Glenn, and K. R. Koedinger. Instruction based on adaptive learning technologies. In *Handbook of Research on Learning and Instruction*, pages 522–560. Routledge, New York, 2 edition, 2016.
- [5] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.
- [6] L. W. Anderson and D. R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman, New York, 2001.
- [7] R. Baker and L. M. Rossi. Assessing the disengaged behaviors of learners. *Design recommendations for intelligent tutoring systems*, 1:153, 2013.
- [8] R. S. Baker. Stupid tutoring systems, intelligent humans. *International journal of artificial intelligence in education*, 26(2):600–614, 2016.
- [9] R. Barnett. *A Will to Learn: Being a Student in an Age of Uncertainty*. Open University Press, Maidenhead, 2007.
- [10] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakçı, and R. Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025.
- [11] A. Birhane. Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2):100205, 2021.
- [12] E. L. Bjork and R. A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher and J. R. Pomerantz, editors, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth Publishers, New York, 2011.
- [13] R. A. Bjork, J. Dunlosky, and N. Kornell. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64(1):417–444, 2013.
- [14] B. S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Longmans Green, New York, 1956.
- [15] C. Borchers and T. Shou. Can large language models match tutoring system adaptivity? A benchmarking study. In *International Conference on Artificial Intelligence in Education*, pages 407–420, Cham, 2025. Springer Nature Switzerland.
- [16] J. D. Bransford, A. L. Brown, and R. R. Cocking. *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington, DC, 2000.
- [17] J. S. Brown, A. Collins, and P. Duguid. Situated cognition and the culture of learning. *1989*, 18(1):32–42, 1989.
- [18] E. Brynjolfsson, D. Li, and L. Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.
- [19] S. Bulathwela, M. Pérez-Ortiz, C. Holloway, M. Cukurova, and J. Shawe-Taylor. Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. *Sustainability*, 16(2):781, 2024.
- [20] S. Bull. There are open learner models about! *IEEE Transactions on Learning Technologies*, 13(2):425–448, 2020.
- [21] S. Bull. There are open learner models about! *IEEE Transactions on Learning Technologies*, 13(2):425–448, 2020.
- [22] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.
- [23] M. Cazzaniga, F. Jaumotte, L. Li, G. Melina, A. J. Panton, C. Pizzinelli, E. J. Rockall, and M. M. Tavares. Gen-AI: Artificial intelligence and the future of work. Staff Discussion Note SDN/2024/001, International Monetary Fund, 2024.

- [24] M. T. Chi and R. Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243, 2014.
- [25] J. Chung, M. Henderson, C. Slade, Y. Liang, N. Pepperell, T. Corbin, J. Walton, A. S. Yu, M. Bearman, S. Buckingham Shum, T. Fawns, T. McCluskey, J. McLean, G. Oberg, A. Seligmann, A. Shibani, A. Bakharia, L.-A. Lim, and K. E. Matthews. The use and usefulness of GenAI in higher education: Student experience and perspectives. *Computers and Education Open*, 7:100347, 2026.
- [26] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.
- [27] B. J. Corbett and J. M. Tangen. Ai tutors vs. tenacious myths: Evidence from personalised dialogue interventions in education. *Computers in Human Behavior*, 175:108828, 2026.
- [28] E. Cosyn, H. Uzun, C. Doble, and J. Matayoshi. A practical perspective on knowledge space theory: Aleks and its data. *Journal of Mathematical Psychology*, 101:102512, 2021.
- [29] A. Darvishi, H. Khosravi, S. Sadiq, D. Gašević, and G. Siemens. Impact of ai assistance on student agency. *Computers & Education*, 210:104967, 2024.
- [30] F. Dell’Acqua, E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Kraye, F. Candelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of Artificial Intelligence on knowledge worker productivity and quality. *Organization Science*, 2026. First released as Harvard Business School Working Paper No. 24-013, September 2023.
- [31] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643, 2019.
- [32] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu. Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies. *Computers & Education*, 227:105224, 2025.
- [33] P. Denny, S. MacNeil, J. Savelka, L. Porter, and A. Luxton-Reilly. Desirable characteristics for ai teaching assistants in programming education. ITiCSE 2024, page 408–414, New York, NY, USA, 2024. Association for Computing Machinery.
- [34] D. Dermeval, R. Paiva, I. I. Bittencourt, J. Vassileva, and D. Borges. Authoring tools for designing intelligent tutoring systems: A systematic review of the literature. *International Journal of Artificial Intelligence in Education*, 28(3):336–384, 2018.
- [35] E. Dillon and colleagues. Microsoft 365 copilot and knowledge worker productivity: Field experiment. Working paper, 2025. Randomised experiment across 56 firms; placeholder metadata.
- [36] S. D’Mello and A. Graesser. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica*, 151:196–207, 2014.
- [37] S. Doroudi, V. Alevan, and E. Brunskill. Where’s the reward? a review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29(4):568–620, 2019.
- [38] H. Drachler and W. Greller. Privacy and analytics: it’s a delicate issue—a checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge (LAK 2016)*, pages 89–98. ACM, 2016.
- [39] Y. Fan et al. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2):489–530, 2024.
- [40] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena. An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.
- [41] W. Gao, Q. Liu, L. Yue, F. Yao, R. Lv, Z. Zhang, H. Wang, and Z. Huang. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23923–23932, 2025.
- [42] A. Ghosh, N. Heffernan, and A. S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2330–2339, 2020.
- [43] A. C. Graesser, N. K. Person, and J. P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522, 1995.
- [44] X. Han, H. Peng, and M. Liu. The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies. *Educational Research Review*, 45:100714, 2025.

- [45] J. Hardy, S. P. Bates, M. M. Casey, K. W. Galloway, R. K. Galloway, A. E. Kay, P. Kirsop, and H. A. McQueen. Student-generated content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*, 36(13):2180–2194, 2014.
- [46] M. Henderson, M. Bearman, J. Chung, T. Fawns, S. Buckingham Shum, K. E. Matthews, and J. de Mello Heredia. Comparing generative ai and teacher feedback: student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, Online: 13 May 2025:1–16, 2025. doi: 10.1080/02602938.2025.2502582.
- [47] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. Buckingham Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, et al. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3):504–526, 2022.
- [48] M. H. Immordino-Yang and A. Damasio. We feel, therefore we learn: The relevance of affective and social neuroscience to education. *Mind, brain, and education*, 1(1):3–10, 2007.
- [49] S. Järvelä, G. Zhao, A. Nguyen, and H. Chen. Hybrid intelligence: Human–ai coevolution and learning. *British Journal of Educational Technology*, 2025.
- [50] F. J.-Y. e. a. Jin. Students’ perceptions of generative ai–powered learning analytics in the feedback process: A feedback literacy perspective. *Journal of Learning Analytics*, 12(1):152–168, 2025.
- [51] M. Kapur. Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, 51(2):289–299, 2016.
- [52] M. Kazemitabaar, R. Ye, X. Wang, A. Z. Henley, P. Denny, M. Craig, and T. Grossman. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery.
- [53] H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y. S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.
- [54] H. Khosravi, G. Demartini, S. Sadiq, and D. Gasevic. Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th international learning analytics and knowledge conference*, pages 32–42, 2021.
- [55] H. Khosravi, P. Denny, S. Moore, and J. Stamper. Learnersourcing in the age of ai: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 5:100151, 2023.
- [56] H. Khosravi, A. Shibani, J. Jovanovic, Z. A. Pardos, and L. Yan. Generative ai and learning analytics: Pushing boundaries, preserving principles. *Journal of Learning Analytics*, 12(1):1–11, 2025.
- [57] O. Klimecki and T. Singer. Empathy from the perspective of social neuroscience. *The Cambridge handbook of human affective neuroscience*, pages 533–550, 2013.
- [58] K. R. Koedinger, P. F. Carvalho, R. Liu, and E. A. McLaughlin. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, 120(13):e2221311120, 2023.
- [59] K. R. Koedinger, A. Corbett, et al. *Cognitive tutors: Technology bringing learning sciences to the classroom*. na, 2006.
- [60] D. A. Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.
- [61] N. Kornell and R. A. Bjork. A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138(4):449–468, 2009.
- [62] J. W. Lai, W. Qiu, M. Thway, L. Zhang, N. B. Jamil, C. L. Su, S. S. Ng, and F. S. Lim. Leveraging process-action epistemic network analysis to illuminate student self-regulated learning with a socratic chatbot. *Journal of Learning Analytics*, 12(1):32–49, 2025.
- [63] H.-Y. Lee, J. Kim, H. Choi, H. Bae, A. Jeong, S. Choi, J.-H. Kim, and C.-E. Kim. Comparing ai chatbot simulation and peer role-play for osce preparation: a pilot randomized controlled trial. *BMC Medical Education*, 25(1):1755, 2025.
- [64] H. Li, T. Xu, C. Zhang, E. Chen, J. Liang, X. Fan, H. Li, J. Tang, and Q. Wen. Bringing generative ai to adaptive learning in education. *arXiv preprint arXiv:2402.14601*, 2024.
- [65] M. Liffiton, B. E. Sheese, J. Savelka, and P. Denny. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, Koli Calling ’23, New York, NY, USA, 2024. Association for Computing Machinery.

- [66] J. Lin, M. Rakovic, D. Lang, D. Gasevic, and G. Chen. Exploring the politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 282–293, New York, NY, USA, 2022. Association for Computing Machinery.
- [67] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918, 2014.
- [68] A. Meyer, D. H. Rose, and D. Gordon. *Universal Design for Learning: Theory and Practice*. CAST Professional Publishing, Wakefield, MA, 2014.
- [69] J. Meyer and R. Land. Threshold concepts and troublesome knowledge: Linkages to ways of thinking and. *Princeton: Citeseer*, 2003.
- [70] F. Miao, W. Holmes, H. Ronghuai, and Z. Hui. *AI and Education: Guidance for Policy-Makers*. UNESCO Publishing, Paris, 2021.
- [71] I. Molenaar. The concept of hybrid human-ai regulation: Exemplifying how to support young learners’ self-regulated learning. *Computers and Education: Artificial Intelligence*, 3:100070, 2022.
- [72] I. Molenaar, S. de Mooij, R. Azevedo, M. Bannert, S. Järvelä, and D. Gašević. Measuring self-regulated learning and the role of ai: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139:107540, 2023.
- [73] I. Molenaar, A. Horvers, and R. S. Baker. Towards hybrid human-system regulation: Understanding children’s support needs in blended classrooms. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 471–480, 2019.
- [74] S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.
- [75] A. Nugent, J. M. Lodge, A. Carroll, R. Bagraith, S. MacMahon, K. Matthews, and P. Sah. *Higher education learning framework: An evidence informed model for university learning*. The University of Queensland, 2019.
- [76] B. D. Nye, A. C. Graesser, and X. Hu. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469, 2014.
- [77] OECD. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. OECD Publishing, Paris, 2026.
- [78] E. Panadero. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology*, 8:422, 2017.
- [79] E. Panadero and S. Järvelä. Socially shared regulation of learning: A review. *European psychologist*, 2015.
- [80] Z. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of educational technology & society*, 17(4):49–64, 2014.
- [81] P. I. Pavlik, L. G. Eglington, and L. M. Harrell-Williams. Logistic knowledge tracing: A constrained framework for learner modeling. *IEEE Transactions on Learning Technologies*, 14(5):624–639, 2021.
- [82] R. Pelánek. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User modeling and user-adapted interaction*, 27(3):313–350, 2017.
- [83] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot, 2023.
- [84] D. N. Perkins and G. Salomon. Transfer of learning. In *International Encyclopedia of Education*. Pergamon Press, Oxford, 2 edition, 1992.
- [85] P. R. Pintrich. The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, 41(4):219–225, 2002.
- [86] S. Pozdniakov, J. Brazil, M. Mohammadi, M. Dollinger, S. Sadiq, and H. Khosravi. Ai-assisted co-creation: Bridging skill gaps in student-generated content. *Journal of Learning Analytics*, 12(1):129–151, 2025.
- [87] S. Pozdniakov, R. Martinez-Maldonado, S. Singh, H. Khosravi, and D. Gašević. Using learning analytics to support teachers. In *Handbook of Artificial Intelligence in Education*, pages 322–349. Edward Elgar Publishing, 2023.
- [88] A. Rafferty, H. Ying, J. Williams, et al. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Journal of Educational Data Mining*, 11(1):47–79, 2019.
- [89] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.

- [90] S. Ramanathan, L.-A. Lim, N. R. Mottaghi, and S. Buckingham Shum. When the prompt becomes the codebook: Grounded prompt engineering (groproe) and its application to belonging analytics, 2025.
- [91] E. F. Risko and S. J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.
- [92] H. L. Roediger and A. C. Butler. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1):20–27, 2011.
- [93] H. L. Roediger and A. C. Butler. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1):20–27, 2011.
- [94] H. L. Roediger and J. D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006.
- [95] A. Scarlatos, N. Liu, J. Lee, R. Baraniuk, and A. Lan. Training llm-based tutors to improve student learning outcomes in dialogues. In *International Conference on Artificial Intelligence in Education*, pages 251–266. Springer, 2025.
- [96] B. Sheese, M. Liffiton, J. Savelka, and P. Denny. Patterns of student help-seeking when using a large language model-powered programming assistant. In *Proceedings of the 26th Australasian Computing Education Conference, ACE '24*, page 49–57, New York, NY, USA, 2024. Association for Computing Machinery.
- [97] L. S. Shulman. Signature pedagogies in the professions. *Daedalus*, 134(3):52–59, 2005.
- [98] A. J. Smith, K. E. Boyer, J. Forbes, S. Heckman, and K. Mayer-Patel. My digital hand: A tool for scaling up one-to-one peer teaching in support of computer science learning. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, SIGCSE '17*, page 549–554, New York, NY, USA, 2017. Association for Computing Machinery.
- [99] N. C. Soderstrom and R. A. Bjork. Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2):176–199, 2015.
- [100] J. Stamper, R. Xiao, and X. Hou. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer, 2024.
- [101] T. Swist, S. Buckingham Shum, and K. N. Gulson. Co-producing aided ethics under lockdown: An empirical study of deliberative democracy in action. *International Journal of Artificial Intelligence in Education*, 34:670–705, 2024.
- [102] K. Topping. Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology*, 1(1):1–17, 2017.
- [103] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik. The influences of emotion on learning and memory. *Frontiers in Psychology*, 8:1454, 2017.
- [104] UNESCO. Guidance for generative ai in education and research. UNESCO Policy Guidance, 2023. ISBN 978-92-3-100615-5.
- [105] K. VanLehn. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265, 2006.
- [106] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.
- [107] K. VanLehn. *Student modeling*. Psychology Press, 2013.
- [108] L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.
- [109] K. Weatherholtz, K. M. Hill, K. DiCerbo, W. Wells, P. Grimaldi, M. Miller-Vedam, C. Hogg, and B. Yamkovenko. Cognitive engagement in genai tutor conversations: At-scale measurement and impact on learning. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 40–48, 2025.
- [110] J. Weidlich, D. Gašević, H. Drachsler, and P. Kirschner. ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning*, 41(4):e70105, 2025.
- [111] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, 1998.
- [112] T. Wheatley, O. Kang, C. Parkinson, and C. E. Looser. From mind perception to mental connection: Synchrony as a mechanism for social understanding. *Social and Personality Psychology Compass*, 6(8):589–606, 2012.

- [113] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave. The role and limits of principles in AI ethics: Towards a focus on tensions. pages 195–200, 2019.
- [114] P. H. Winne and A. F. Hadwin. nstudy: Tracing and supporting self-regulated learning in the internet. In *International handbook of metacognition and learning technologies*, pages 293–308. Springer, 2013.
- [115] P. H. Winne and N. E. Perry. Measuring self-regulated learning. In *Handbook of self-regulation*, pages 531–566. Elsevier, 2000.
- [116] B. Woolf, W. Bursleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.
- [117] Z. Xu, K. Wijekumar, G. Ramirez, X. Hu, and R. Irey. The effectiveness of intelligent tutoring systems on K-12 students’ reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50(6):3119–3137, 2019.
- [118] L. Yan, S. Greiff, J. M. Lodge, and D. Gašević. Distinguishing performance gains from learning when using generative ai. *Nature Reviews Psychology*, 4:435–436, 2025.
- [119] C. Zhai, S. Wibowo, and L. D. Li. The effects of over-reliance on ai dialogue systems on students’ cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.
- [120] B. J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory into practice*, 41(2):64–70, 2002.