AI Hub for Education

A FORWARD-LOOKING RESEARCH AGENDA FOR AI AND LEARNING

Building the Evidence Base for Genuine K-12 Transformation

Carly D. Robinson
Chris Agnew
Susanna Loeb

November 2026

Stanford | SCALE Initiative
Accelerator for Learning

Introduction

Artificial intelligence is becoming part of students' lives, and education systems are making consequential decisions about its use. AI's effects on learning and development remain uncertain. Research has already documented learning losses from some uses of AI. Education decision makers and product developers, as well as researchers, policy makers and philanthropy, can influence that trajectory through the decisions that they make, the tools that they build, the evidence that they generate, and the uses that they support or limit. This agenda sets out the evidence needed to guide those decisions and expand what students can become and what educators and schools can do for every student.

This agenda is intended to guide research investments by government agencies, foundations, and research teams seeking to generate evidence that can shape policy, practice, and technology development in education. It asks under what conditions AI should and should not be used to support students and educators and, where it helps, what educational systems must change to make it accessible and useful. A complementary agenda from CRPE takes up the adjacent question of how policymakers, system leaders, and funders can use that evidence to guide transformation.

School leaders are setting procurement, governance, staffing, and accountability policies that will lock in patterns for years, and the choices made in coming years will shape students' learning experiences for far longer. The evidence base is not keeping pace. A recent review identified more than 1,600 studies relevant to AI in K-12 education. Only a small fraction provide strong causal evidence, and very few examine U.S. K-12 settings at scale. Timely research can help shape adoption toward approaches that expand learning opportunities and narrow disparities in access and learning.

Meeting this challenge also requires educational research itself to change. The traditional process, in which a single study unfolds over several years of proposals, funding, piloting, and full-scale data collection, will often produce findings that are obsolete by the time they arrive. A central task for the coming decade is to build the capacity to learn from AI adoption while it is still underway. Doing so will require accelerated research cycles, including informative descriptive research, shorter-term A/B tests, and validated short-term measures that serve as proxies for the long-term outcomes that matter.

AI's place in education should be earned, not assumed

This agenda takes seriously the question of where AI should not be used at all—where developmental, intellectual, or well-being considerations make human connection and judgment irreplaceable. The case for adopting AI should rest on evidence that it expands what students can become, what educators can do, and what schools can achieve for every student. Therefore, a finding that an AI approach does not work, or works no better than the non-AI approach, is as valuable to the field as a finding that it does work. Research must be designed to detect the full range of outcomes transparently, and communicate them to the field so that leaders have the capacity to act on these findings while the opportunity is still open.

This agenda rests on four principles that define what counts as progress and how evidence should be built. The principles are the most durable commitments: they apply to any use of AI in education, and they will hold as tools change and specific questions are answered. The agenda then applies these principles to three research bets. The bets examine whether AI can develop the full range of students' capacities, improve instruction through changes in educators' roles, and enable models of schooling that improve student development and expand educational opportunity. Each research bet introduces claims to be tested, not assumptions to be defended.

Four guiding principles: What counts as progress and how to build the evidence

Most research on AI and education asks a narrow question: does this tool improve or hurt outcomes? This agenda asks whether and how AI can strengthen students' long-term development, and how education systems can evolve to support it. Much of the existing research evaluates AI against familiar metrics, treats existing school structures as fixed, and largely sidesteps whether current education is preparing students for the world they will actually inhabit. Transformation requires rethinking both what counts as success and how public systems are designed to pursue it. Following from the premise that AI's place in education must be earned, these four principles are the agenda's foundation.

Principle 1: Improvement means building a broad range of students' capacities

To thrive socially, economically, and in civic life, young people need academic knowledge, but also a range of cognitive, social, emotional, and self-regulatory capacities that will shape their long-term trajectories. Capacities such as critical thinking, discernment, adaptation, creativity, and empathy will be important in a world where AI changes how information is produced, evaluated, and used. Yet most current research, and most accountability systems, evaluate AI against narrow performance metrics such as test scores, course completion, and short-term skill gains.

This agenda aims for a higher standard. AI should be evaluated based on whether it strengthens students' developing capacities across all of these dimensions and leads to balanced and equitable development. It should also be evaluated on where it does harm, or displaces forms of human connection and judgment that students need. Research designs therefore need to detect effects on sensemaking, persistence, curiosity, collaboration, and self-regulation as well as on a standardized test. These capacities, shown in Table 1 (reproduced from Gonzalez et al., 2026), span interrelated dimensions of development that evolve over time and are shaped by learning environments. Together they represent the range of capacities that prepare students to thrive in higher education, the workforce, and civic life.

Table 1. Capacities for Thriving

Skills and Knowledge	Affects
Academic Knowledge and Skills	**Motivation**
Academic knowledge is factual and conceptual understanding in a targeted subject domain, and academic skills are learned procedures and strategies for applying subject-specific knowledge.	Engagement driven by enjoyment or interest in learning (intrinsic motivation) or by external rewards, expectations, or obligations (extrinsic motivation).
Higher-Order Thinking Skills	**Interest and Curiosity**
Abilities in problem-solving, question asking, creativity, innovation, and critical thinking. These skills are analysis, synthesis, and generative thinking.	Desire to explore, ask questions, and resolve gaps in understanding.
Social Skills	**Belonging and Interpersonal Connection**
Collaboration, persuasion, and communication with peers or adults.	Learners' perception that they are valued, accepted, and legitimate members of the learning environment.
Metacognition, Self-Regulation, and Adaptability	**Self-Efficacy, Mindset, and Self-Concept**
Planning, monitoring, and adjusting one's own learning.	Beliefs about one's own ability to learn and grow through effort.
Autonomy Skills	**Management of Context-Specific Anxiety, Boredom, and Frustration**
The capacity to make independent choices.	Negative emotions towards the subject or tool that can hinder persistence and learning focus.

Note - These academic and thinking skills are drawn from the P21 Framework (Partnership for 21st Century Skills, 2009) and the OECD Learning Compass 2030 (OECD, 2019); social skills from CASEL Core Competencies (Collaborative for Academic, Social, and Emotional Learning, 2020) and the P21 Framework (Partnership for 21st Century Skills, 2009); metacognition and self-regulation skills from Farrington's framework of Non-Cognitive Factors (Farrington et al., 2012) and CASEL Core Competencies (CASEL, 2020); adaptability and curiosity from the OECD Learning Compass 2030 (OECD, 2019); autonomy skills and motivation from Self-Determination Theory (Ryan & Deci, 2000); belonging, interpersonal connection, self-efficacy, and mindset from Farrington's framework of Non-Cognitive Factors (Farrington et al., 2012); and context-specific emotions management from CASEL Core Competencies (Collaborative for Academic, Social, and Emotional Learning, 2020)

Principle 2: Data, measurement, and evaluation are essential infrastructure

A key constraint on AI's potential in education is the current state of the underlying data, the existing measures the field has to work with, and the misaligned incentives for evaluation. Education systems need data infrastructure that supports both the development of effective AI tools and the ability to learn systematically from their implementation over time. Effective AI systems require high-quality training data, and education currently lacks two foundational kinds: data on what excellent instruction looks like and how students actually respond to different instructional approaches. Observational records of skilled teaching are sparse, rarely annotated, and seldom available in formats usable for model development. Detailed data on student learning trajectories (e.g., how learners respond to different explanations, where they get stuck, and what misconceptions they carry) remains locked in proprietary systems or simply uncollected. Without this data, AI models trained for educational use inherit limitations from whatever datasets were available during development, which typically reflect neither the range of effective instructional practice nor the diversity of student learning experiences. Building the data infrastructure needed to support better educational AI, including standards for collection, annotation protocols, and privacy-protecting sharing agreements, is itself a research priority.

Measurement is foundational to this agenda. If effects on students' capacities cannot be detected reliably, it becomes difficult to evaluate whether AI-supported approaches are helping or harming students. The capacities most vulnerable to AI-related harm, including agency, persistence, and professional judgment, are also the hardest to measure. This difficulty creates a recurring risk that harms will outpace the field's ability to detect them. Each bet below identifies its own version of this measurement lag risk. Investments in developing and validating measures of these capacities, including measures derived from learning interactions and behavioral data, are therefore a prerequisite for much of the research that follows.

However, measurement alone does not build an evidence base. The field also needs credible evaluation to determine whether AI-supported approaches actually improve outcomes, for whom, under what conditions, and at what cost. This requires research designs that can identify causal effects, distinguish short-term signals from durable outcomes, and examine implementation and context.

The challenge is not just methodological. The incentives to make data available, release measures, and generate rigorous, independent evidence are often weak. Developers, and the schools and districts that adopt their products, may have little incentive to subject them to evaluations that could produce negative findings. Schools and districts may lack the resources or capacity to conduct rigorous evaluations of their own, while legitimate concerns about student privacy can constrain data sharing. Researchers, too, often face stronger incentives to produce novel, standalone studies than to replicate findings, use shared measures or contribute to a cumulative evidence base. Addressing these barriers requires infrastructure that allows evidence to accumulate across studies and settings. When each team develops its own outcomes, findings cannot be compared or pooled; when studies draw on common, validated measures and use comparable designs across samples, subjects, and

grade levels, results can accumulate across sites into a coherent body of evidence. Building that evidence base is a collective task that calls for coordinated investment in shared data and measures, independent evaluation, and multi-site studies. This type of shared infrastructure has enabled rapid progress in other fields.

Principle 3: Progress involves scaling what works and inventing what's next

AI is reshaping what students need to know and be able to do, transforming how information is generated and evaluated, how work is organized, and how people participate in civic and economic life. Yet the core delivery model of public education has changed little, with age-graded classrooms, fixed schedules, subject silos, and standardized accountability systems still defining teaching and learning. This agenda treats school structure as an object of redesign, and it therefore operates across two horizons at once.

The first horizon is to scale what works. Across every domain there are approaches with real evidence behind them that have begun to scale, where the open questions are about reach, cost, and implementation. The second is to invent what's next, including approaches that could be transformative but are still being built, where the forms are unsettled, the measures are immature, and the evidence is thin. Both kinds of work appear across every part of the agenda. Table 2 provides illustrative questions of each type of work within core domains.

Table 2. Examples of near-term and transformational work across core domains

	Scale what works (near-term)	Invent what's next (transformational)
Learning experiences	Can AI extend high-impact tutoring and personalized support to students who have never had access?	How might AI enable new forms of instruction that build the full range of student capacities?
Educators' work	Can AI offload routine tasks and help educators use known-effective practices more consistently?	How might AI make fundamentally redesigned educator roles and staffing models workable?
School systems	Can AI optimize efficiency within existing school structures?	How might schooling itself need to evolve, and can AI make new models feasible?

The field often treats these as competing priorities. This agenda puts the two in conversation with one another instead. They carry different risk profiles and timelines: the near-term work is tractable

and measurable, while the transformational work is harder to pin down. Because AI evolves faster than long-term outcomes can be observed, both kinds of work depend on developing validated short-term proxies for those outcomes, so that research can iterate on rapidly changing tools without waiting years to learn whether they worked.

Investing only in approaches with established evidence may limit opportunities to explore new models that AI could make possible. Investing only in the not-yet-scaled ignores the needs students and educators have today. A natural objection is that structural and organizational research should wait until basic efficacy evidence for AI tools is established, so that systemic redesign is not built on an unproven foundation.

This agenda rejects that sequencing for two reasons. First, education systems are already making binding structural decisions, procurement contracts, staffing models, governance arrangements, before tool-level efficacy evidence exists. A sequential approach would arrive after those decisions are locked in rather than before. Second, tool-level efficacy and structural conditions are not separable in the way sequencing assumes. Whether a given AI tool works depends on the organizational conditions and educator roles it is embedded in, so structural research generates the context needed to interpret efficacy findings correctly, not a premature bet on unproven tools.

Progress on the near-term horizon builds the partnerships, measurement infrastructure, and credibility that make the transformational work possible. The transformational work keeps the near-term work honest about what schooling is ultimately for.

Principle 4: What works should reach students least well served first

Educational technology has a history of benefiting advantaged students more and widening gaps in the process. The sharpest version of this risk with AI is relational, not just technological. AI-mediated instruction could become a substitute for human relationship for students with the least resources, while better-resourced students retain both human connection and AI support. Reaching students with the least access first does not mean testing unproven tools on them. These students have the most to lose from approaches that do not work, and AI should add to the human support they receive, not replace it. It means that when approaches show evidence of benefit, these students should get them first rather than last. It also means research should be designed from the outset to learn whether approaches work for these students, not only for those who are easiest to reach. This agenda embeds the question of whether AI reduces or widens access gaps directly into each research question it asks.

Closing gaps requires more than new tools or better evidence. It requires changes in how resources are allocated, how roles are defined, and who holds authority over instructional decisions. Those changes can generate resistance from educators, administrators, families, and institutions with stakes in existing arrangements. That resistance is sometimes reasonable. Research going forward should therefore include attention to the political and organizational dynamics of AI-enabled change, including how stakeholder attitudes vary and shift over time, how power is redistributed when AI changes what educators do and who has access to data, and what change management and advocacy approaches build genuine buy-in rather than surface compliance. These are not peripheral

concerns. They are among the most practically important questions a research agenda focused on transformation can ask.

Three research bets: Hypotheses about where research on AI could make a difference for students

This section applies the four principles above to three promising research bets about how AI could address persistent challenges in education. Research will examine whether hypothesized benefits emerge, for whom, and under what conditions, alongside the harms and costs. Each bet includes questions about the design of tools, their use in schools, and the organizational conditions that shape their effects. The research is designed to produce findings that can guide decisions about which approaches to develop, expand, revise, limit, or discontinue.

The bets span the three levels through which AI may influence education: **students' learning experiences**, **educators' work**, and **the systems in which both are embedded**. Each is grounded in a persistent challenge education has not solved. Each bet, if supported by evidence, has substantial implications for policy, practice, and the students who have historically been least well served.

How the bets are organized

Each bet follows the same structure: the problem it addresses, how AI might help, the potential harms and costs, and a set of priority research questions.

The bets are not sequential. They move outward from students to educators to the systems in which both are embedded, but they should be pursued at the same time, because each depends on the others. A tool that strengthens learning depends on educators equipped to use it well, and a redesigned educator role depends on school structures that can sustain it. Research on all three should proceed in parallel.

Studies designed around specific products or current model capabilities risk obsolescence before they are published. Research focused on mechanisms and enabling conditions asks a different question: what AI-supported experiences produce which outcomes, for which students, and under what conditions. Findings framed this way can inform decisions across successive generations of AI tools.

Four complementary forms of evidence

Four complementary forms of research are needed to guide educational decision-making:

1. **Landscape**: describing current patterns of adoption and use
2. **Impact**: understanding impacts on students, educators, and systems
3. **Enabling conditions**: identifying enabling conditions for responsible and equitable implementation
4. **Evolution**: exploring how schooling itself may evolve as new contexts and technologies reshape how knowledge is generated, evaluated, and applied

These forms of evidence answer different questions and require different methods, and progress in one area strengthens the others, creating a more complete picture of how AI is affecting students, educators, and educational systems. A balanced research portfolio includes all four. To make this concrete, each priority research question below is tagged with the evidence type it primarily generates: landscape, impact, enabling conditions, or evolution.

Across the three research bets, AI's relationship to a given task takes one of three forms. It can *substitute* for a task a person currently performs, *augment* a person's ability to perform that task while leaving them in control, or *target* scarce human and technological resources toward the students, moments, or decisions that most need them. Substitution and augmentation describe how AI changes learning experiences and educator work task by task; targeting describes how AI reallocates scarce resources at the level of the system. Several priority research questions below draw on these categories directly.

The research questions within each bet are intended to guide coordinated portfolios, not isolated studies. Of the many questions that this agenda could prompt, not all are equally urgent. Across the three research bets, highest priority goes to questions that (1) affect large numbers of students, (2) influence decisions already being made by schools and policymakers, and (3) generate evidence that remains useful as AI technologies evolve. Illustrative study designs that show how each bet could be pursued in practice appear in the appendix.

Adjacent lines of research beyond this agenda

This agenda concentrates on the evidence needed to guide decisions about AI in education as adoption unfolds. Several adjacent lines of research are essential to that work but fall outside its scope and should advance in parallel, including:

- *The learning science it builds on.* How people learn, what drives durable understanding, and which supports produce lasting outcomes are foundational questions this agenda draws on

rather than re-answers; progress on AI in education depends on that knowledge base continuing to grow.

- *What good teaching becomes in an AI-shaped world.* Defining effective pedagogy when AI is part of the instructional environment is a parallel inquiry that will shape how the questions here are answered.
- *Legal, regulatory, and cybersecurity analysis.* Specialist work on legal compliance, the design of regulatory frameworks, and technical protections for student data falls outside this agenda's main focus. Findings from that parallel work can inform the design, adoption, and evaluation of educational AI. The agenda includes empirical research on how AI affects students and educators, including how privacy requirements, procurement practices, vendor incentives and governance shape implementation and outcomes.

An evolving agenda

This agenda is meant to be revisited. Regular review can update the priority questions and research bets as evidence accumulates and educational uses for AI change. Findings from adjacent lines of research may also change the agenda's scope. The four principles can continue to guide the choice of tools, uses, outcomes, methods, and populations studied.

Research Bet 1: How can AI help create learning experiences that strengthen the full range of student capacities?

The problem

The promise of personalized learning has been central to educational technology for decades. In practice, most "personalized" systems adjust the pace or difficulty of standardized content. True personalization requires detecting where a student is, diagnosing why they are struggling, and matching them with experiences and supports suited to their specific needs—a task that is cognitively and logistically difficult at scale.

Personalization has also traditionally been aimed at a narrow target. Systems optimize for what they can easily measure, like test scores, course completion, and discrete skills. The capacities that matter most for long-term thriving are far harder to assess. When success is defined by these metrics, the instruction often narrows to match, and the harder-to-measure capacities are left underdeveloped.

High-quality instruction and individualized support have always been distributed unequally. Students in under-resourced schools and districts, students with disabilities, English learners, and students from low-income families consistently have less access to the kinds of support that drive long-term success. Decades of research have identified effective forms of support, yet access to those supports remains constrained by resources, staffing, and delivery capacity. AI may make some forms of individualized support more feasible to deliver at scale than previous technologies allowed, and the bet is that it can become an expected standard for all students, not a privilege of the most advantaged.

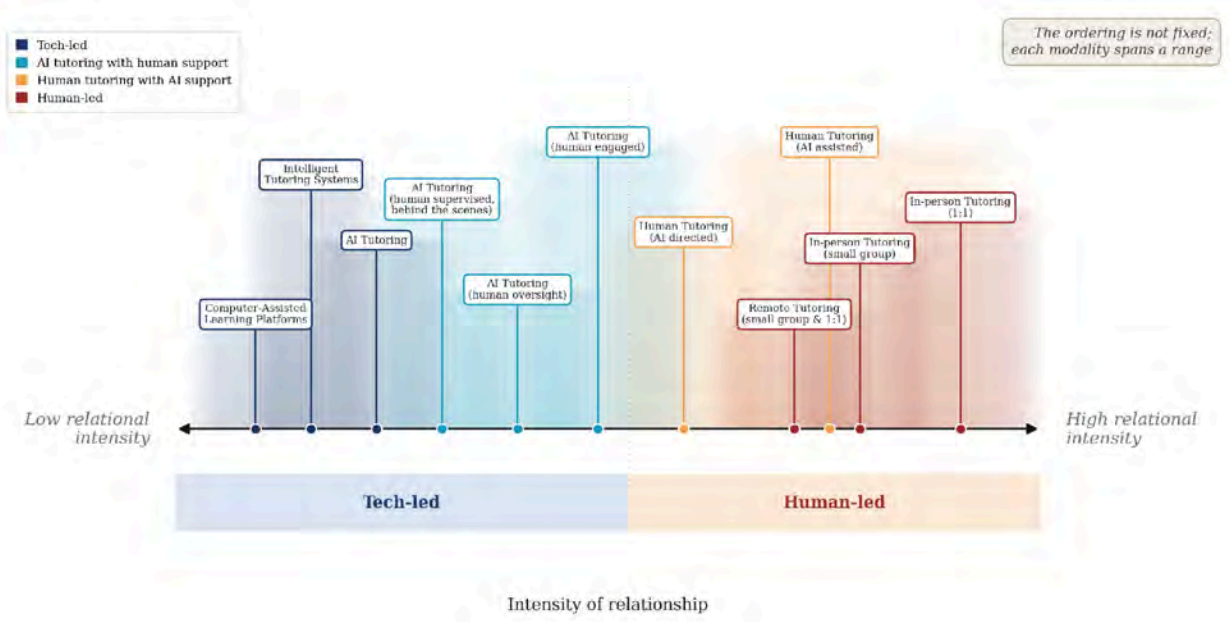
Why now

The value of personalized support is well established. What is new is its reach, and the possibility of creating learning experiences that previous technologies could not support. High-impact tutoring can produce large learning gains, yet access has always been constrained by the availability of skilled adults. AI changes that constraint by making responsive, individualized support possible without a one-to-one supply of expert educators. The open question is whether AI can deliver comparable benefits at population scale. Answering that question depends on measurement systems that can detect not only academic gains, but also effects on motivation, agency, persistence, and other capacities that personalized support is intended to strengthen.

AI also changes what it is possible to measure. AI-supported learning environments generate continuous behavioral evidence, including moment-to-moment signals about how students engage, where they slow down, how they respond to difficulty, and how they work with others. Traditional assessments do not capture these signals. AI opens the possibility of measuring capacities that have always mattered but been largely invisible, such as persistence, curiosity, sensemaking, and the quality of collaborative work. Developing and validating these measures is foundational to the research agenda. AI-supported tools are already being adopted at scale; in many cases they are becoming embedded in daily routines before the measurement infrastructure needed to evaluate them exists. Patterns of overreliance, reduced student agency, or differential access may take root before they are visible in outcome data, which is precisely why investing early in measurement is both an intellectual and a practical priority.

One useful way to organize this work is to recognize that these instructional experiences vary in both the intensity of the human relationship involved and the degree of AI involvement. Figure 1 maps personalized-instruction models along a spectrum of relational intensity, from technology-led models such as computer-assisted learning platforms to human-led models such as one-on-one in-person tutoring. The ordering is not fixed, and each modality spans a range. Just as schools currently vary the type and intensity of interventions students receive based on their needs, the appropriate balance of AI and human involvement may also differ across students, learning goals, and contexts. Rather than identifying a single optimal configuration, the goal is to understand what mix of technological and relational support works best for whom, when, and for what purpose. The value of the spectrum is that it makes visible the wide range of instructional configurations that combine human relationships and AI support. Schools are already choosing among these configurations, often without strong evidence about which are most effective for which students and purposes.

Figure 1. Personalized instruction: On a spectrum of relational intensity



Caption: A theoretical depiction of models of personalized instruction by relational intensity, from tech-led (low relational intensity) to human-led (high relational intensity). Each modality spans a range, not a point, as indicated by the shaded band, so the ordering is not fixed. For instance, an AI-assisted human tutor—with AI handling logistics so the human invests more in the relationship—can reach or even exceed the relational intensity of in-person tutoring. The vertical position of labels is arbitrary and used only to keep labels legible.

The risk

If AI-supported experiences raise short-term performance without building durable capacities, or quietly erode persistence, agency, independent reasoning, and social connection through overreliance, then scaling them would entrench shallow learning at population scale, and the costs would fall hardest on students with the fewest alternatives. This measurement lag risk is named in Principle 2, where the capacities most at risk of erosion (persistence, agency, independent reasoning, belonging and interpersonal connection) are also the hardest to measure, so harms could spread before they appear in outcome data. Distinguishing productive support from harmful substitution will be paramount.

Priority research questions

- ***Measuring what matters (enabling conditions):*** **How can education systems measure the capacities that matter in ways that are valid, reliable, sensitive to change, and practical to use at scale?** Research can develop and validate measures of learners' capacities associated with long-term outcomes; test whether behavior- and interaction-based measures function equivalently across student groups and contexts; and identify measures that may systematically misrepresent or disadvantage particular students.

- ***Affecting durable learning and skills (impact):*** **Do AI-supported learning experiences produce durable gains in the capacities in Table 1, for which students, and under what conditions do they fail to outperform non-AI alternatives?** Relevant outcomes include academic knowledge, higher-order thinking, social skills, and metacognition. Research should be designed to detect no difference or a negative effect, not only a positive one, and should distinguish durable learning from surface-level gains, examining risks such as cognitive offloading, reduced autonomy, and overreliance on automated support. Research can also identify which approaches most effectively serve students who have historically been underserved, and which do not.
- ***Thriving in an AI world (evolution):*** **What capacities matter in an AI-shaped world, and how do learning experiences cultivate them?** Research can identify how the capacities students need to thrive may change as AI evolves and reshapes how people learn, work, and participate in society. This includes examining which capacities, such as discernment, creativity, collaboration, and ethical judgment, become more important, and how learning environments can cultivate them through inquiry, problem solving, and authentic work.
- ***Understanding the role of relationships and human support (impact):*** **What does the human relationship contribute to motivation, persistence, and learning, and under what conditions does AI augment or erode those contributions?** Research can identify the aspects of human connection that drive engagement and persistence, the moments at which they are most important, and whether AI extends or crowds out relational support. This line of work also explores how different configurations of relational and technological intensity shape student learning outcomes, with implications for costs and feasibility.
- ***Prioritizing implementation and context (enabling conditions):*** **What implementation conditions sustain student motivation and engagement in AI-supported learning over time?** Research can examine how pacing, feedback design, novelty, and the balance of AI and human interaction affect whether early motivational gains hold up over a full school year, and what instructional and staffing conditions are needed to sustain them at scale.
- ***Reaching students with the least access (landscape):*** **Is AI-enabled personalized support actually reaching students in under-resourced schools, students with disabilities, English learners, and students from low-income families, or is it following the adoption pattern of prior ed-tech and reaching better-resourced students first?** Research can track uptake and implementation quality across school and district resource levels, identify the barriers (cost, infrastructure, staffing, training) that determine which schools adopt AI-supported personalization first, and test interventions designed to accelerate reach into under-resourced settings.

Research Bet 2: How can AI help enable fundamentally different educator roles that improve instruction and student support?

The problem

Educators work under conditions that make consistent, high-quality instruction difficult to sustain. High caseloads, limited feedback loops, insufficient planning time, and professional learning systems that are episodic and disconnected from real-time classroom decisions can all constrain educators.

They spend significant time on content delivery, routine feedback, grading, and administrative tasks, leaving limited capacity for the relationship-building, targeted intervention, and deep instructional design that most shape student development. The result is a persistent gap: decades of research have produced robust evidence about effective instruction, yet that knowledge rarely reaches daily practice at scale. The consequences often fall unevenly: students in under-resourced schools are least likely to experience consistent, high-quality implementation of what the research says works. AI creates an opportunity to close this gap by embedding evidence-based practice directly into the instructional environment and freeing educators for the work that most benefits from human expertise. Realizing the opportunity depends on whether systems deliberately reinvest the time AI saves into higher-value work rather than treating it simply as efficiency gains.

Why now

What is newly possible is not just doing today's job better but redefining the job itself. As AI absorbs tasks that have always constrained how teaching is organized, staffing and role configurations that were never feasible become possible. A team-based staffing model illustrates this possibility: one experienced teacher could oversee several paraprofessionals or less-experienced teachers across a larger group of students, with AI handling real-time progress monitoring, flagging students who need the lead teacher's attention, and surfacing instructional guidance to the paraprofessionals as they work. This configuration was previously difficult because no one had the bandwidth to track every student's progress closely enough to know who needed the lead teacher's expertise at a given moment.

By taking on routine work and surfacing evidence-based guidance at the moment instructional decisions are made, AI can change what educators spend their time and expertise on. That capacity also helps close the research-to-practice gap that has long resisted previous solutions, largely because evidence-based guidance is dispersed during episodic professional learning disconnected from practice. Educators are not unaware that more evidence exists than gets used; the issue is that current teaching conditions leave little room for adoption and sustained change. The same lever could reach other structural problems that stem from scarce time and support: uneven instructional quality, inconsistent implementation, and weak feedback loops.

Skilled, well-supported educators provide a benchmark for evaluating AI-supported instruction. Research can examine whether and how AI improves instruction where resources and support are limited, and whether it helps educators devote more time to relationships, targeted intervention, and instructional design. Studies can identify which combinations of educator expertise, AI support, and organizational conditions improve student outcomes.

The risk

If time saved is reabsorbed by administrative demands rather than redirected to high-value work, or if embedding guidance in the moment narrows teaching and erodes professional judgment, then redesigned roles could deskill educators and weaken relationships without improving outcomes. A related risk is overreliance on the educator side: teachers may come to lean on AI in ways that erode their own expertise, and training pipelines may prepare a workforce that assumes AI will always be

available, leaving educators unable to teach effectively when it is not. This measurement lag risk, named in Principle 2, applies here too. Deskilling and eroded professional judgment may not show up in outcome data until after they have taken hold. Because staffing structures, contracts, and preparation pipelines are slow to reverse, if this bet proves wrong, it would be costly to undo.

Priority research questions

- ***Disentangling educator and AI configurations (impact):*** **Under what conditions do configurations of educator roles and AI supports match, exceed, or fail to reach the outcomes of skilled, well-resourced human instruction alone?** Research can examine which educator tasks AI substitutes for entirely, which it augments while leaving the educator in control, and which it helps target, directing attention or resources to the students or moments that most need them, and how these three modes affect outcomes, cost, and professional judgment differently.
- ***Promoting take-up of evidence-based practices (enabling conditions):*** **How can AI strengthen educator judgment and increase the consistent use of evidence-based practices?** Research can examine how AI supports implementation of practices such as the science of reading and high-quality mathematics instruction, including through AI-enabled formative assessment tools that surface real-time, actionable feedback to teachers and tutors on individual students' understanding, how educators incorporate AI-generated recommendations into instructional decisions, and how AI-generated insights can be shared appropriately with families and community partners to support learning beyond school.
- ***Increasing efficiency and redeploying effort (impact):*** **How does AI reshape educators' use of time, attention and expertise, and what organizational conditions determine whether those changes improve practice?** Research can examine whether efficiency gains are redirected toward relationship-building, targeted intervention, instructional planning, and professional collaboration, or absorbed by administrative demands and other competing priorities. Additionally, research can test whether AI primarily raises the floor in under-resourced or inconsistent teaching contexts, amplifies already-strong educators, or both, and how to pair AI with skilled educators so it removes low-value work without displacing the judgment and relationships that define excellent teaching.
- ***Identifying the hallmarks of an educator (evolution):*** **What knowledge, skills, and dispositions do educators need as AI reshapes the work of teaching?** Research can examine the knowledge, skills, and dispositions educators need to work effectively alongside AI, as well as how preparation programs and professional learning systems adapt to redesigned roles.
- ***Understanding workforce changes (evolution):*** **How does AI-driven role redesign affect the teaching workforce, including recruitment, retention, deskilling, and who chooses to enter teaching?** Research can examine effects on recruitment, retention, job quality, compensation structures, specialization, career progression, and attrition, as well as how changing educator roles influence who chooses to enter and remain in the profession.
- ***Reaching under-resourced schools and classrooms (landscape):*** **Are AI-enabled changes to educator roles and staffing reaching the schools and classrooms with the least consistent access to high-quality instruction, or are they concentrating in already well-resourced**

systems? Research can compare adoption patterns, implementation supports, and role redesign across schools that vary in resources and instructional quality, and identify what conditions (funding, leadership, professional learning capacity) determine whether under-resourced schools can adopt redesigned educator roles at all.

Research Bet 3: How can AI help enable new models of schooling that improve student development and expand educational opportunity?

The problem

For many students, the current model of schooling does not work. Age-graded classrooms, fixed schedules, standardized pacing, and one-teacher-per-classroom staffing leave large numbers of students under-challenged, under-supported, or disengaged, and decades of reform within that model have not changed it much. Most research on AI in education focuses on whether a particular tool improves outcomes within this structure. The more consequential question is whether AI now makes it possible to change the structure itself, re-organizing how students spend their time, how educators are deployed, what students are expected to learn, how support is delivered, and how resources are allocated. AI may expand the range of options available across each of these dimensions to build schooling that works for students it currently fails.

Concretely, AI can change what schools can do with their scarcest resources including adult expertise, student time, information, and money. It can make it feasible to group and regroup students by need rather than age, to give every student individualized practice and feedback without a one-to-one supply of adults, and to put real-time information in front of the people making instructional and staffing decisions. These capabilities introduce a set of models long proposed but rarely achievable at scale such as competency-based progression, flexible and extended learning time, team-based staffing, and integrated multi-tiered support. It remains an open question whether these proposed models produce different outcomes than schools that add AI to otherwise unchanged structures. Few studies examine AI as part of broader organizational redesign. This bet treats school design itself as an object of study.

The early evidence for redesigned models is thin, and some of it is discouraging. The most visible attempts to build AI-centered schools have so far produced underwhelming or unverified results, and much of what circulates is marketing rather than evidence. Early results warrant caution in both directions, because deep redesign takes time to work. New models need iteration, capacity-building, and time for students and educators to adapt before their effects become visible. Judging them on a first year or two risks discarding approaches that were immature rather than ineffective. This pattern is what the agenda's two-horizon logic predicts: the transformational work is where the evidence is thinnest, the failure rate highest, and the timelines longest.

The appropriate response is neither to assume redesigned models work nor to dismiss them on early returns. It is to study them rigorously over realistic timelines, distinguishing models that are not yet working from those that will not, while holding open the real possibility that some do no better, or

worse, than incremental uses of AI. Replacing unverified claims with credible evidence about which redesigns help students, and which do not, is itself a central contribution of this bet.

Whether promising models reach students depends on more than their design. As prior waves of educational technology showed, the effects of new tools depend on the systems into which they are introduced. Public education systems operate within governance structures, accountability frameworks, procurement processes, professional norms, and political constraints that shape which innovations reach students, how they are implemented, whether they persist beyond pilot phases, and who ultimately benefits. Yet, most school systems currently lack coherent guidance for AI integration.

School systems also shape the kinds of educational models that become possible. Procurement criteria, evidence standards, accountability metrics, interoperability requirements, and transparency expectations influence what developers build and bring to market. These signals affect whether AI tools support new approaches to organizing learning, deploying expertise, and allocating resources, or reinforce existing structures. When those signals are weak, fragmented, or misaligned with long-term student development, the resulting tools may optimize for short-term adoption at the expense of long-term educational value. This problem is not limited to school redesign. Principle 2 discusses the field-wide incentive problem in evidence generation. Within Bet 3 specifically, schools that have adopted redesigned models face the disincentive that independent evaluation could expose disappointing results or proprietary methods. As a result, evidence on whether redesign works is undersupplied precisely where it is most needed.

Why now

AI makes it possible to improve existing school structures while also exploring fundamentally different models of schooling. Whether those models produce better outcomes at scale remains largely unknown, and where early evidence exists, it is mixed. Research has the greatest opportunity to influence practice before patterns become established. This is also a period in which market expectations are still being formed. The evidence standards, procurement criteria, accountability measures, and interoperability requirements that education systems adopt today will help determine what developers build tomorrow. Research can help ensure that those signals are aligned with student development, educational equity, and long-term learning goals rather than short-term adoption pressures. Whether new models of schooling improve student outcomes depends on their design, implementation and the governance and policy conditions under which they operate.

The risk

Because system redesign reshapes schooling itself, its mistakes get built into the structure rather than layered on top of it. A redesign could fail to improve outcomes, or improve them only for already-advantaged students while leaving others with thinner, AI-mediated support. It could also embed new harms regardless of its effect on achievement, building inequities, surveillance, or vendor dependence in schools. Deep redesign takes years to show effects. This delay is the measurement lag risk named in Principle 2 in its most acute form. The evidence needed to catch a failing model may not arrive until the model is already embedded in contracts, staffing, and routines.

System redesign requires substantial investments of time, money, and organizational capacity. Pursuing ineffective models can divert attention from more promising approaches and make existing disparities more difficult to address. Above all, the cost falls on students, who pass through school only once and cannot recover years spent inside a model that does not work.

Priority research questions

The questions below refer to models such as those introduced above, including competency-based progression, flexible and extended learning time, team-based staffing, and integrated multi-tiered support, whether adopted individually or as part of a fully redesigned school.

- ***Assessing outcomes and cost (impact):*** **Under what conditions do AI-enabled school redesigns improve outcomes and cost-effectiveness beyond what incremental AI use within existing structures achieves, and under what conditions do they fail to do so or underperform?** Research can examine outcomes related to student development, learning opportunities, educator workload, instructional quality, and cost-effectiveness, as well as how savings are distributed or reinvested. Research can also identify which design features transfer across different contexts and student populations.
- ***Documenting emerging models (landscape):*** **How are school models, staffing structures, accountability systems, and decision-making processes changing in response to AI?** Research can document emerging models, variation across contexts, and patterns associated with coherence, responsiveness, and equity.
- **Evolving as models sustain and scale (*evolution*): How do AI-enabled school models change in form as they move from pilot to sustained implementation, and does what makes a model effective at small scale still hold as it scales?** Research can examine how governance arrangements, staffing models, and training practices adapt or are abandoned over time, and whether models that persist end up resembling their original design or evolve into something substantially different.
- ***Aligning incentives with equitable design (enabling conditions):*** **How do funding formulas, accountability systems, procurement practices, and governance structures shape the development of effective and equitable AI-enabled school models?** Research can examine the incentives these arrangements create and how those incentives influence educational opportunities for different groups of students.
- ***Shaping the educational AI market (impact):*** **How do evidence standards, procurement criteria, interoperability requirements, accountability metrics, and transparency expectations influence the educational AI market?** Research can examine how these signals affect the tools that are developed, adopted, and scaled, as well as the governance and political dynamics that influence which signals are most tractable to change.
- ***Redefining governance and professional roles (evolution):*** **How do governance structures and professional roles change as AI reshapes what school systems need to do well?** Research can examine how data safeguards, interoperability systems, professional learning, and leadership

structures change as AI becomes embedded in school operations, and how systems avoid entrenching harmful surveillance or vendor lock-in as those structures shift.

- ***Reaching under-resourced systems (landscape):* Are AI-enabled school models being adopted and sustained in under-resourced districts, or are they emerging first, or only, in well-resourced systems?** Research can compare which districts adopt redesigned models, what capacity (funding, staffing, governance stability) is required to sustain them, and whether models developed in well-resourced settings transfer to districts with fewer resources.

Conclusion

Four guiding principles focus this agenda on building students' capacities; investing in data, measurement, and evaluation; extending effective practices and developing new models; and bringing effective approaches to students least well served. The three research bets apply these principles to AI's role in students' learning experiences, educators' work, and the organization of schooling. Studies that connect these areas can reveal how changes at each level shape outcomes.

Advancing this agenda depends on coordinated investment in shared data and measures, independent evaluation, longitudinal and multi-site studies, and partnerships among researchers, developers, and education systems. Research can identify benefits, harms, and costs across students and settings. Findings can inform the design of tools, educators' decisions, and the policies that determine which approaches reach students.

The ambition is to expand what students can become and what educators and schools can help them achieve. Research can help identify the learning experiences that build durable capacities, the ways of organizing educators' work that strengthen instruction and relationships, and the school models that expand opportunities for students least well served today. Building that evidence gives researchers, developers, and funders, alongside educators, families, and policymakers, a consequential role in shaping education's future.

Appendix: Illustrative study designs

The priority questions above are intended to guide coordinated research portfolios, not isolated studies or tool evaluations. The examples below illustrate how the agenda's central questions could be translated into concrete, rigorous research. The examples span the three research bets and the cross-cutting infrastructure needed to build a cumulative evidence base. They are illustrative rather than prescriptive.

Study 1: Validating measures of student capacities (Cross-cutting infrastructure)

A measurement study could develop and validate proxy measures of the capacities in Table 1, including persistence, agency, sensemaking, and self-regulation, using behavioral and interaction data already generated by AI-supported learning environments. Researchers would build candidate measures from signals such as help-seeking patterns, response latency, self-correction, and engagement with difficult material, drawing on students already enrolled in AI-supported tutoring or practice programs across multiple sites.

The study would test predictive validity by following students after their AI-supported learning ends, comparing early behavioral signals against later performance on unfamiliar tasks completed independently, without AI support. This would establish which short-term behavioral measures reliably predict durable, independent learning, and how far in advance they can do so. The study would also test measurement invariance, examining whether the same behavioral signal carries the same meaning across different student populations, or whether it systematically over- or under-represents the capacities of particular groups.

This work is a prerequisite for other studies in this agenda, not a standalone evaluation. Studies 2 through 4 rely on measures of persistence, sensemaking, and self-regulation to assess equity and durability. The validity of their conclusions depends on those measures functioning as intended across the student populations those studies sample.

Equity considerations

A measure that predicts outcomes well for one group but not another would misrepresent the capacities of the group for which it fails, and any study relying on that measure would draw the wrong conclusion about whether AI is helping or harming those students. The study would specifically test measurement invariance across English learners, students with disabilities, and students in under-resourced schools, since these are the groups most likely to interact with AI-supported tools in patterns that differ from the samples used to build behavioral measures. Where a measure functions differently across groups, the study would document how it fails, so later studies can correct for the bias or avoid relying on that measure for the populations where it is unreliable.

Study 2: AI tutoring, measurement, and durable learning (Research Bet 1)

A multi-site study could examine whether the short-term gains produced by AI-supported tutoring persist after students no longer have access to AI support. Students would be randomly assigned to AI-supported tutoring, human tutoring, or business-as-usual instruction during a defined instructional unit. Researchers would measure performance immediately following the intervention and again several months later, including tasks that require students to transfer their learning to unfamiliar problems and perform independently without AI assistance.

The study would pair conventional assessments of academic learning with measures of conceptual understanding, transfer, sensemaking, persistence, and self-regulation, including measures derived from students' interactions with AI-supported learning environments. Students could be followed beyond the immediate intervention to test which short-term measures and behavioral signals predict later independent performance and durable learning. Using common measures and protocols across participating sites would also test whether these measures produce comparable results across different students, schools, and instructional contexts.

Equity considerations

The study would recruit samples spanning school resource contexts and student characteristics, including English learners and students with disabilities, with sufficient power to examine heterogeneous effects. Researchers would test both whether AI-supported tutoring produces different effects across student groups and whether the measures themselves function differently across those groups. A central question would be whether apparent differences in outcomes reflect differences in learning or limitations in how particular capacities are being measured.

Study 3: Human and AI configurations for personalized instruction (Research Bet 1)

A randomized controlled trial could compare how different configurations of human and AI support affect learning, development, and student experience. Configurations might include AI tutoring with no regular human support; AI tutoring with weekly teacher check-ins; human tutoring supported by AI diagnostics and session planning; and small-group tutoring with AI-supported practice between sessions. Students would be randomly assigned to available models within participating schools, with outcomes measured at the end of the semester and again six months later.

The study would examine effects on achievement, conceptual understanding, persistence, motivation, and students' sense of agency. It would also track how students use the AI tools, both in and between sessions, and whether patterns of reliance on AI predict later independent performance. Rather than assuming a single optimal balance of human and technological support, the study would test whether the effectiveness of each configuration varies by students' prior achievement, learning needs, and access to other forms of support. Implementation analysis would include an analysis of dosage, continuity, cost, and how schools match different configurations to

students' needs over time. A key question would be whether AI expands access to higher-intensity support or instead creates a two-tier system in which students with fewer resources receive more heavily AI-mediated experiences.

Equity considerations

High-relational instructional experiences, including one-on-one tutoring, small-group mentorship, and close teacher-student relationships, have always been unequally distributed. The study would examine whether AI-enabled configurations can bring higher relational intensity to students who currently receive the least, and whether the benefits of human relationship in learning can be partially extended through AI support or require direct human presence. Findings would have direct implications for how districts allocate human and technological resources across schools with different starting points.

The study would be designed from the outset to recruit samples spanning income levels, school resource contexts, and student characteristics, including English learners and students with disabilities, with sufficient power to detect heterogeneous effects.

Study 4: AI-supported collaborative learning (Research Bets 1 & 2)

A randomized study could compare classes using three versions of a project: conventional project-based learning; project-based learning with an AI system that provides feedback and prompts to individual students; and project-based learning in which the AI supports the teacher in monitoring group progress but does not interact directly with students.

Researchers would measure knowledge, quality of student reasoning, collaboration, persistence, and ability to explain and defend conclusions independently of the AI. Classroom observations and interaction data would examine whether AI changes who participates, how groups divide cognitive work, and whether students become more or less reliant on the system over time.

The research would test whether these environments strengthen capacities such as sensemaking, discernment, creativity, collaboration, persistence, and independent reasoning, rather than relying primarily on short-term academic performance. It would also examine whether particular forms of AI support strengthen these capacities or instead encourage cognitive offloading, reduced autonomy, or superficial engagement. The study could compare models that use AI primarily as a scaffold with models in which AI takes on a larger share of the cognitive work.

Equity considerations

Collaborative and project-based learning has historically been concentrated in higher-resourced schools. Research would examine whether AI scaffolding makes these models more accessible and effective in under-resourced settings, and whether AI tools support or undermine participation for students who have been historically marginalized within group work, including students with disabilities, English learners, and students with lower academic confidence.

Study 5: Redesigning educator work around AI (Research Bet 2)

A school-level study could compare schools using AI to support selected planning, feedback, and progress-monitoring tasks with schools using existing workflows. In participating treatment schools, teachers would use AI for specified routine functions, with an explicit plan to redeploy a portion of time saved toward small-group instruction, student conferencing, and collaborative planning.

Researchers would measure student outcomes alongside teacher time use, instructional practice, workload, professional judgment, and retention. The study would test whether efficiency gains actually produce more high-value educator work or are absorbed by other demands. A longer-term component could examine whether continued AI use strengthens educators' expertise or contributes to deskilling and overreliance.

Equity considerations

The study would examine whether AI improves the consistency of instructional quality in schools with high turnover or limited access to professional learning, and whether changes in educator roles affect different categories of educators or students differently.

Study 6: AI-enabled formative assessment and teacher/tutor-facing feedback (Research Bets 1 & 2)

A school- or classroom-level study could compare teachers and tutors using an AI-enabled formative assessment tool, one that analyzes student responses, work samples, or practice problems in real time and surfaces diagnostic feedback about individual students' understanding, with teachers and tutors using conventional formative assessment practices such as periodic quizzes, exit tickets, or informal checks for understanding.

Researchers would measure whether AI-generated diagnostic feedback changes what educators notice about student understanding, how quickly they identify students struggling with a specific concept, and whether they adjust instruction, grouping, or pacing in response. Student-level outcomes would include content mastery, the persistence of misconceptions over time, and whether early diagnostic accuracy translates into more effective subsequent instruction. The study would also examine educator trust and use: whether teachers and tutors act on the tool's recommendations, override them, or ignore them, and under what conditions each response occurs.

This study addresses the near-term horizon by testing AI as a way to extend established formative assessment practices. The open research questions concern whether AI-generated feedback improves instruction and student learning, how educators use it, and what conditions support effective use at scale.

Equity considerations

The study would examine whether AI-generated diagnostic feedback is equally accurate across student populations, including English learners and students with disabilities, since formative assessment tools trained primarily on general education response patterns may misread partial understanding, language-related answer patterns, or non-standard problem-solving approaches as misconceptions. It would also examine whether educators in under-resourced schools, who often have the least planning time to act on additional information, are able to use the feedback the tool provides, or whether the tool primarily benefits classrooms that already have the staffing and time to respond to it.

Study 7: Team-based staffing models for instruction (Research Bet 2)

A school-level pilot could test a team-based staffing model in which one experienced lead teacher oversees several paraprofessionals or less-experienced teachers across a larger group of students, with an AI system providing real-time progress monitoring, flagging students who need the lead teacher's direct attention, and surfacing instructional guidance to paraprofessionals as they work with students. Participating schools would be compared with schools using conventional one-teacher-per-classroom staffing at similar grade levels and subject areas.

Researchers would measure student learning outcomes, the frequency and appropriateness of the lead teacher's interventions, whether flagged students actually received timely attention, and the quality of instruction delivered by paraprofessionals under AI guidance. The study would also examine feasibility: how the model affects staffing costs, whether it can be sustained beyond an initial pilot, and how much training paraprofessionals need before they can act effectively on AI-surfaced guidance. A central question would be whether the AI's monitoring and flagging genuinely extends the lead teacher's expertise to more students, or whether errors and gaps in the AI's judgment go uncaught without a second expert available to verify them.

This study addresses the longer-term horizon by testing whether AI-supported monitoring and instructional guidance can make team-based staffing more effective and sustainable. The open questions concern whether the model works at all, and under what conditions, rather than how to extend something already known to work.

Equity considerations

The study would examine whether team-based staffing is a realistic option for under-resourced schools, which may have the hardest time recruiting an experienced lead teacher willing to take on this expanded role, or whether the model only becomes viable in schools that already have stronger staffing and leadership capacity. It would also examine whether students in the paraprofessional-led portion of instruction receive a different quality of experience than students who happen to be assigned more of the lead teacher's direct time, and whether that assignment pattern correlates with

student need, disability status, or English learner status in ways that could concentrate the least experienced instruction on the students who need the most.

Study 8: AI-enabled scheduling and staffing (Research Bet 3)

At the system level, a set of schools could pilot an AI-assisted scheduling model that dynamically groups students for intervention and enrichment rather than relying entirely on fixed support periods. The system could use student performance and attendance information to help determine when students receive small-group instruction, AI-supported practice, and individual or small-group teacher support.

Researchers would compare these schools with schools using conventional scheduling and examine effects on student learning and broader capacities, access to high-quality instructional time, educator workload, and cost. The study would test whether AI allows schools to allocate scarce human expertise more effectively or simply shifts more students toward AI-mediated support.

Equity considerations

A core question would be whether flexible, AI-enabled scheduling expands access to individualized supports for students who currently receive the least, or whether schools reproduce existing hierarchies by allocating high-quality human time to already-advantaged students while routing others to AI-mediated instruction. Research would also examine staffing implications for paraprofessionals and support staff, whose roles and job security are most directly affected by AI-enabled restructuring.

Study 9: Governance and decision-making in AI adoption (Research Bet 3)

A qualitative or mixed-methods study could follow districts as they make major AI procurement and implementation decisions over two years. Researchers would document what evidence districts request from vendors, what vendors provide, whether districts conduct independent evaluation, how privacy and data-sharing requirements affect evaluation, and how decisions change as early evidence emerges.

The study could compare districts using explicit evidence requirements in procurement with districts using more conventional processes. The research would examine which products are adopted, what evaluation practices are established, whether vendors change their evidence practices, and whether districts develop stronger capacity to learn from implementation. The study would also examine what happens when early findings are mixed or negative. This research could surface systemic barriers and enabling conditions and clarify how policy and governance reforms might better align AI integration with student capacity development.

Equity considerations

The study would specifically examine how AI adoption decisions play out differently across districts that vary in geography, capacity, resources, and the populations they serve. It would investigate whether AI procurement and governance processes embed equity requirements, whether independent community voices shape decisions alongside vendors and administrators, and whether accountability systems create incentives to serve historically underserved students or to optimize for populations already performing well.

Study 10: AI support for students with learning differences (Research Bets 1, 2 & 3)

This study examines two sides of AI-supported special education: the family-facing side and the educator-and-system side.

Family-facing component. A study could examine how AI tools that surface timely, interpretable information about student learning trajectories change family engagement, advocacy, and the quality of home-school communication. For families of students with disabilities in particular, navigating special education processes, IEP development, and service coordination is time- and knowledge-intensive in ways that disadvantage families with less social capital, familiarity with systems, or time to engage. Research would investigate whether AI-generated summaries of student progress, rights-relevant information, and meeting preparation tools change how families participate in decisions about their children's education, and whether those changes translate into better-matched services, stronger IEP goals, and improved student outcomes.

Educator-and-system component. A parallel study could examine whether AI support for IEP development and progress monitoring reduces administrative burden, improves the quality of documentation, and frees special educators to spend more time on instruction and direct student support. A key question is whether AI tools help special education staff act on data more quickly, or simply add a new layer of documentation without changing practice.

Equity considerations

Family-facing tools risk widening the very advocacy gaps they could close. Families with more formal education, stronger English proficiency, reliable connectivity, and time to engage are best positioned to act on AI-generated summaries. Families with less social capital may benefit least. The study would examine three questions. Do these tools narrow or widen participation gaps across language, literacy, immigration status, and digital access? Are translations and plain-language summaries accurate and genuinely usable across the languages and reading levels families actually have? Is the information surfaced equally valid for students from different groups? The study would also attend to trust and surveillance. Families who are wary of institutions, including immigrant and historically marginalized families, may experience these tools as support or as monitoring, since that perception shapes whether the tools reach the families who stand to gain the most.

On the educator side, students with disabilities are among the most consistently underserved by educational AI tools, which are typically designed and validated on general education populations and often fail to account for diverse learning profiles, assistive technology needs, and communication differences. The study would examine whether the benefits of reduced administrative burden are distributed evenly across special education programs, or concentrated in programs with more specialist staffing and stronger technology infrastructure to begin with.

Acknowledgements: We are grateful to the Spencer Foundation for its support of writing this research agenda. We thank the CRPE team for their thought partnership and for advancing the ideas in this agenda through their complementary work. We also thank Overdeck Family Foundation, the Walton Family Foundation, and the Stanford Accelerator for Learning for their continued support. We appreciate the many people who provided feedback on earlier drafts, including Emma Brunskill, Thomas Dee, Dora Demszky, Isabelle Hau, Guilherme Lichand, and Philip Oreopoulos. Finally, we thank the entire SCALE Initiative team for their thoughtful contributions and for pushing this work forward.