Built for Learning

Lessons From Emerging Artificial Intelligence Solutions and Approaches

By Marisa Mission, Michelle Croft, and Amy Chen Kulesa

JANUARY 2026

Bellwether

CONTENTS

Introduction

The integration of artificial intelligence (AI) into ed tech tools has raised myriad questions about how such advanced technology can both ease burdens for students and teachers and facilitate deep learning. Bellwether's prior work has explored this tension by diving into how AI could amplify productive struggle and how to measure the impact of AI-powered ed tech tools.[1] This series subsequently showcases those concepts in practice: Drawing on interviews with education leaders and ed tech founders from more than 20 organizations nationwide, this brief identifies common design trade-offs, implementation challenges, and emerging approaches. Five accompanying case studies dive deep into specific organizations, illustrating how these broader themes can play out across contexts and use cases. Across the series, key themes include:

1. **Balancing cognitive support with productive struggle**: Thoughtful organizations make deliberate choices about when AI should ease burdens versus when difficulty drives learning.

2. **Centering appropriate roles for teachers versus technology**: Most organizations center educators and high-quality instructional materials (HQIM), focusing on using technology to augment, rather than automate, existing resources and processes.

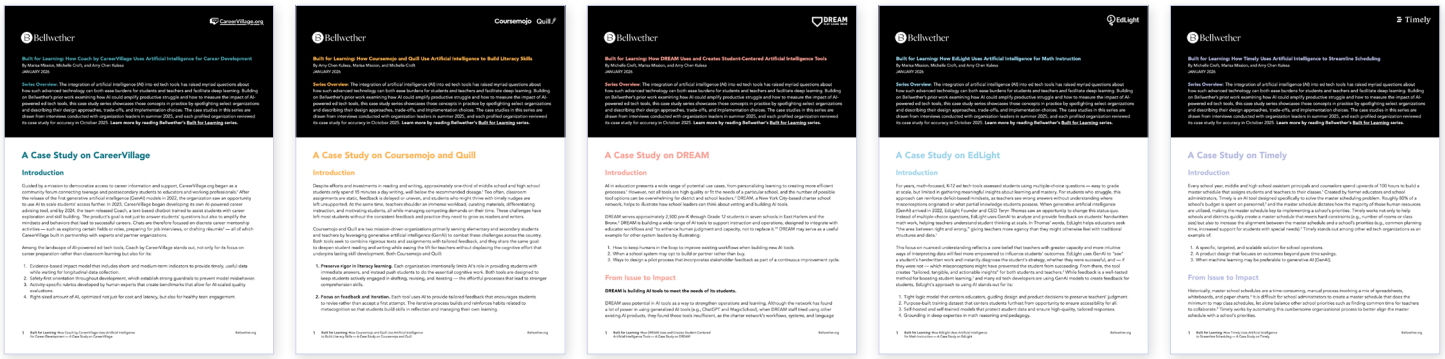
3. **Navigating measurement challenges to evaluate the medium- and long-term impact of AI-powered tools**: Despite significant challenges — including cost, rapid product iteration, and limited longitudinal data — thoughtful developers are working to measure meaningful learning outcomes rather than just usage.

4. **Building infrastructure that reflects both practical and pedagogical considerations**: Technical decisions about tool architecture, data handling, model selection, and more carry implications for privacy, accessibility, and learning experiences for students.

5. **Identifying sustainable business models that balance pedagogical commitments and market realities**: As the market of AI-powered ed tech grows and frontier AI models advance, smaller organizations face questions of pricing, scale, and competition from general-purpose AI tools.

The six organizations profiled span a diverse range of settings, from literacy, writing, and mathematics instruction to career-connected learning and school operations. Some organizations serve elementary students, while others target high school or adult learners. What unites them is a commitment to centering pedagogy in their technical decisions and grappling seriously with questions of impact.

CareerVillage **Coursemojo and Quill** **DREAM** **EdLight** **Timely**



While each case study explores specific organizations' design choices, it is likely too early to point at any one tool or practice as "the best." Instead, the goal of this series is to help education leaders better understand the design choices, technical approaches, and implementation considerations that distinguish thoughtful AI integration from rushed adoption. Practitioners and educators can use these themes to ask sharper questions of vendors about how their tools will use AI to best serve their specific students. Ed tech developers might find new ideas or approaches for addressing common challenges. And policymakers and funders will be able to better assess whether AI tools align with their educational priorities, and what infrastructure or ecosystem conditions are critical for effective, sustainable implementation.

Technology in education, whether AI-powered or not, should always center student well-being and learning. However, the novelty and technological power of generative AI (GenAI) can obfuscate its true positive or negative impact. This overview brief and the accompanying case studies attempt to pull back the curtain on AI-powered ed tech and offer a synthesized set of learnings about the choices developers face when designing their tools.

Designing AI Tools for Impact

Whether an AI-powered ed tech tool meaningfully deepens learning — as opposed to just adding more technology in the classroom — starts with how the tool is designed to work. GenAI models might carry out the work, but the tool itself must be grounded in thoughtful decisions and theories of action that center student outcomes.[2] When designing for impact, ed tech leaders interviewed for this brief elevated four key considerations: integrating productive struggle, amplifying good teachers, building on HQIM, and measuring student outcomes.

Organizations vary in how they use AI to support productive struggle.

The ed tech leaders interviewed understood the dangers of letting students offload cognitive processes to AI; most have made deliberate choices about how their tools support productive struggle. At minimum, all developers maintain guardrails preventing students from immediately, effortlessly getting answers. But beyond these guardrails, different organizations employ a variety of strategies to increase students' productive struggle, including calibrating problem difficulty to individual students, providing feedback that encourages students to keep thinking, and cultivating students' motivation and growth mindset.

Training AI outputs to cultivate both learning and motivation can be tricky, though. One organization noticed that when a lesson plan builds on previous activities and a student is stuck on an earlier concept, it was more appropriate for the tool to show the answer, allowing the student to keep up with the class rather than falling behind and disengaging entirely. (The tool would still make a note of the student's prior progress to ensure they could get necessary support from a teacher later.) This example illustrates a more nuanced understanding of productive struggle: Effective tools make intentional choices about when to scaffold and when to push, recognizing that the goal is sustained, effortful engagement rather than struggle for its own sake.

Why does productive struggle matter for AI-powered ed tech? In Productive Struggle, Bellwether investigates how productive struggle — or "the process of engaging with challenging tasks or problems that require effort, critical thinking, and persistence to solve" — deepens learning by enhancing cognitive elements such as information processing, sustained focus, motivation, and metacognition. Incorporating productive struggle in AI-powered products is one way ed tech designers can ensure their product meaningfully contributes to students' learning.

Most developers view teachers as essential partners in the learning process, not as inefficiencies to automate away.

Across all of the interviews, organizations positioned teachers — not technology — as the primary drivers of learning. Many ed tech founders were teachers themselves, and they frequently cited this background when explaining their design philosophies. For some tools, this manifests through giving teachers control over how much students see of the AI's feedback or suggestions, preserving teacher autonomy in the classroom. Other organizations also see their tools as opportunities for embedded teacher professional development, especially for newer teachers. For example, as they see how the tool responds to student work, novice educators may be exposed to different instructional strategies they can adopt in their own practice.

Most tools are trying to ease administrative burdens for educators. This includes tracking and using data more effectively, synthesizing trends from student work, and suggesting next steps for instruction. Many aim to reduce the administrative burden so teachers can focus on actual teaching. One developer described their approach as "making it easier for teachers by imitating effective teachers' moves," while another said, "We want to help teachers do their best work more sustainably." These developers are trying to integrate into existing teacher workflows rather than reducing teachers' roles or requiring them to adopt entirely new processes.

Most — but not all — ed tech leaders reject the idea of AI-generated materials, instead choosing to build on existing high-quality curricula.

The relationship between AI tools and HQIM proved to be one of the most thoughtful areas of discussion among ed tech leaders. For most, AI represents an opportunity to help teachers use HQIM more effectively,

but a minority of organizations didn't see these materials as essential. As one leader pointed out, "A good teacher can take a mediocre lesson and scaffold it appropriately, but a lower-quality teacher still might not be able to deliver for students even using an HQIM lesson." For these developers, AI is a helpful tool for generating activities based on curriculum standards and teacher input. Some also noted that for certain extracurricular or career-related courses, well-established "high-quality" curricula simply don't exist, so AI-generated materials can fill a genuine gap.

The majority of ed tech leaders interviewed, however, agreed that HQIM are critical for learning and refuse to let AI generate curricular materials from scratch, citing concerns with whether the output would truly be high quality. One organization believes that if it had not built its tool on a foundation of HQIM, its partner schools would not have achieved the gains in student achievement that they have: "Because the materials are high quality, it becomes more of an attention and engagement issue so that students are doing more thinking, more reading, and more writing." Another developer questioned the purpose of using AI-generated materials: "There's already great curriculum out there, and many school districts have already invested in it, so we don't feel the need to reinvent [it] … the more interesting problem is, what can you do with students' responses to make it meaningful for both the student and teacher?"

The hope is that AI tools might help teachers adapt HQIM for their specific contexts while maintaining the quality of the underlying materials. To that end, almost all organizations have made or are working toward making their platforms flexible enough to integrate with multiple curricula. Some are using in-house or contracted specialists to write activities or lesson plans that work alongside existing curricula. Others are partnering directly with curriculum providers — such as Carnegie Learning, Fishtank, Great Minds, or Illustrative Mathematics — to embed materials directly from the providers. These partnerships have the added benefit of addressing copyright concerns, since the curriculum comes directly from publishers.

Ultimately, there may be an opportunity for AI to help curriculum specialists fill gaps or connect disparate resources. The field, however, appears to be coalescing around the view that AI should serve as a bridge between high-quality curriculum and classroom realities rather than duplicate work that has already been done.

Despite challenges, developers are moving past engagement metrics to measure real teaching and learning gains.

A key indicator of an organization's thoughtfulness around using AI was its ability to articulate — and measure — how its tool would impact student and/or teacher behavior to drive student outcomes. Despite resource constraints and the rapidly evolving nature of GenAI, most organizations are conducting impact research through pilots or early partnerships, with a few pursuing formal studies through academic partnerships. These efforts are beginning to elevate certain short- and medium-term metrics that connect AI-powered tool usage and student outcomes.

Common short-term outcomes focus on student or teacher perceptions and/or mindsets. Simple feedback mechanisms like thumb ratings or single-question, Likert scale surveys at the end of a session can indicate whether users found the AI outputs helpful for learning. Mindset indicators, such as the number of times a student attempted an assignment and how their score changes across attempts, can reflect persistence and willingness to engage with challenge. And one organization reviews the quality of student responses as a measure of student effort and meaningful interaction with the instructional materials.

Medium-term measures typically include more robust student and teacher surveys such as Net Promoter Scores, teacher evaluations, and qualitative feedback gathered through user interviews. Some organizations customize outcome measures to district priorities such as specific instructional goals. Finally, the emerging formal studies have analyzed traditional measures of student achievement (e.g., standardized test scores) as well as data districts already have (e.g., teacher retention rates).

The challenges of measuring AI-powered tools' true impact are significant and suggest a role for ecosystem leaders to support not just tool development, but also the evaluation infrastructure and rapid-cycle research needed to understand AI tools' impact. One developer noted, "We need to shorten the cycle time of responding to new things," suggesting "micro randomized controlled trials" rather than traditional large-scale studies. Supporting these more agile evaluation approaches can help the field build evidence while AI capabilities continue to evolve.

What should measuring AI-powered ed tech tools look like? Many ed tech tools use easily measured outputs such as number of users, frequency of logins, or hours spent (or saved) using a tool. But these metrics do not necessarily reflect whether the tool truly improved instruction or fostered deeper learning. In Measuring Artificial Intelligence in Education, Bellwether offers a road map for leaders and developers interested in selecting meaningful metrics and designing evaluations that capture how AI tools improve student outcomes.

Building High-Quality AI Tools

Thoughtful design decisions must be paired with thoughtful technical implementation. Across the organizations interviewed, all are making deliberate choices about infrastructure, model selection, accessibility, and quality assurance to reflect both pedagogical priorities and practical constraints, including costs and ability to scale services. This section highlights common considerations for ed tech developers, different technical options organizations have chosen, how they measure AI performance, and where ed tech tools have decided to forego using GenAI.

Developers are navigating complex infrastructure trade-offs specific to AI tools.

Building education technology with AI requires navigating a complex set of technical and operational decisions, chief among them accuracy, latency, and cost. Accuracy is the backbone of many tools; without high-quality responses, a tool's purpose is moot. Latency, or speed, matters significantly for efficiency and real-time use cases. If a tool takes too long to respond, teachers will not use it during class time, and students will lose focus. And the fixed and startup costs for AI-powered tools can run into the millions of dollars. These include talent costs — especially for AI-knowledgeable engineers — as well as sales, distribution, and customer service (or implementation) costs.

Variable costs present their own challenges. Per-token pricing[3] and capacity constraints make it difficult for organizations to predict their costs ahead of time, especially as advances in AI models lead to differential pricing. Desirable features incur additional costs: For example, giving a tool persistent memory (where the AI remembers context from previous interactions) requires resending previous responses, increasing token usage (and costs) significantly. Longer conversations therefore can get "exponentially more expensive" as one interviewee pointed out, and can run up against context window limitations. All of these factors make it difficult to estimate unit costs per user, which in turn complicates scaling strategies.

Privacy considerations shape infrastructure choices as well. Generally, most organizations follow typical de-identification practices, such as masking personally identifiable information or not collecting it at all. Some use AI-powered moderation systems to flag instances where students or teachers might be sharing personal information inadvertently. But a small subset of organizations choose to use local processing, where AI models are offline or self-hosted, to ensure greater privacy. This infrastructure offers minimal to no external data sharing, as well as better handling of copyrighted content, but it can increase costs significantly.

Finally, organizations are considering accessibility and inclusivity across multiple dimensions, including connectivity, AI model training, user interface design, and method of interaction. This means questioning assumptions such as 1-1 device ratios or consistent internet access. Organizations are also examining the data used to train AI models, its effects on the resulting AI outputs, and whether both the training data and outputs are truly representative of the students being served. And some organizations are purposefully keeping students with disabilities in mind when designing to maximize accessibility. Other organizations are considering the implications of the data used to train AI models, its effects on the resulting AI outputs, and whether both the training data and outputs are truly representative of the students being served. And some organizations are purposefully keeping students with disabilities in mind when designing to maximize accessibility. This has included ensuring interoperability with screen readers or accessibility controllers; designing user interfaces with larger buttons to minimize distractions; and leveraging multimodal interactions (e.g., using text, audio, and video) throughout a tool. One organization redesigned its entire user manual to ensure that it was accessible through videos, captions, and translations.

Optimizing for quality, latency, costs, privacy, and accessibility is a complex process. Each of those considerations has implications both for technical infrastructure and impact, and there does not seem to be a universally optimal approach. However, AI-powered tools also offer a variety of technical options that can guide developers as they consider trade-offs.

Technical design choices reflect different priorities and constraints.

While every ed tech tool is different, two common decisions are central to every AI-powered tool: model architecture and methods of refining AI responses. For each decision, developers have several choices, resulting in different ways to build similar tools. No one set of choices seems to be "best," as each carries a different set of trade-offs, but organizations can customize their approach to their priorities.

Model Architecture

The pace of advancement in frontier models means that developers have options not just from different providers (e.g., OpenAI, Anthropic, Google), but also through different versions (e.g., Gemini 2.0 versus Gemini 2.5 Flash). While using just one model is the simplest to engineer, almost no organization chose this "single-shot" approach. Instead, most tools rely on multiple models. Some find better quality from one model but quickly run into capacity constraints that force them to switch. Others emphasized that the redundancy of having multiple models minimizes chances of downtime or outage. Several organizations found that different models seem to have better responses for different tasks; a "pipeline" of models therefore allows each model to specialize in tasks, especially if there is audio or visual input. And a major consideration is whether smaller models (e.g., OpenAI's "mini" series) can produce just as high-quality responses using fewer tokens, thereby lowering variable costs.

Another key consideration for model architecture is whether to "build your own" — also known as taking a model offline or self-hosting. This involves downloading an open-source large language model

(LLM) so that queries to the model run privately, rather than using a proprietary LLM that runs on the provider's equipment. For example, Meta's Llama models can be disconnected from Meta, downloaded, and run on a personal computer, whereas OpenAI's ChatGPT-5 must always run on OpenAI's servers.[i] A major upside to this architecture is greater control over privacy and model training. However, self-hosting can become prohibitively expensive as users and queries scale; in fact, most organizations do not self-host, especially as proprietary models are constantly updated. For specialized applications, however, a minority of organizations see the extra costs as worth it to ensure high-quality responses and protect students' data privacy.

Refining AI Responses

Beyond selecting models, organizations can employ additional strategies to ensure that AI responses are accurate and high quality. A common one is leveraging system prompts, which dictate how a model should act when responding to a user's query. More extensive system prompts mean more detailed instructions for the AI model, which lead to more refined responses, especially when the instructions include diverse examples. In a similar vein, context engineering involves compressing and structuring information about a user's history, preferences, and needs to account for factors like a teacher's instructional priorities. Both prompting and context engineering strategies give AI models more information to use when generating responses so that the outputs are customized for quality or personalization.

In contrast, some organizations use retrieval-augmented generation (RAG), which limits the amount of data a model references.[4] In doing so, RAG forces models to respond using vetted information rather than relying solely on their general training. This helps tools perform well within specific domains by grounding responses in custom or educational content rather than the model's broader (and sometimes unreliable) knowledge base. A small number of organizations also leverage web search capabilities to incorporate current information into their responses. This can be particularly valuable for subjects where information changes rapidly or for answering questions that require up-to-date data. RAG and web search integration curate the information AI models reference when generating responses to reduce the risk of including irrelevant or inaccurate information (often referred to as hallucinations).

Lastly, specialized inputs or outputs may require a process called fine-tuning, which involves training general-purpose models on curated data to adapt them for specific use cases.[5] For example, some LLMs struggle to solve math problems, but these models can be fine-tuned to excel at mathematical reasoning (and in turn, math instruction). Essentially, fine-tuning helps models perform better on specific tasks, but it requires substantial amounts of high-quality, domain-specific data as well as additional engineering complexity. As a result, very few organizations choose to go through the process, but it remains a valuable strategy for developers looking to ensure high-quality AI responses for specific niches.

Ensuring high-quality AI outputs requires ongoing testing and refinement.

The inherent "black box"[6] nature of AI means that the technology will not always behave as expected, so developers are often testing AI responses using benchmarks and evaluations. Benchmarking involves setting quality standards for AI responses;[7] often these are rubrics created with experts in content or subject areas. The benchmarks are then used to grade model responses in evaluations. If the model scores poorly against the benchmarks, then the developers use that information to modify or refine their setup until it scores satisfactorily. To scale these evaluations, some organizations use LLMs to do the scoring, then validate those scores through inter-rater reliability tests with human graders.

Consistent testing is critical for a few reasons. One reason is that evaluations can ensure AI model responses are safe for students. General-purpose LLMs have recently come under fire for their inclination to flatter or validate user thinking, even when that thinking is harmful.[8] By including safety as a dimension of testing

i *As of November 2025.*

and evaluation, most ed tech tools avoid similar risks. Another reason is that by using the same benchmarks and evaluations, developers can compare results from different configurations of model architecture and refinement techniques to find the best one for an organization's particular needs. And similarly, being able to compare models facilitates cost-cutting. One developer said, "We benchmark responses based on the more expensive model, then run it with the cheaper model and assess [it] against the benchmark," which allows them to find the highest-quality responses for that specific use case at the cheapest, most sustainable rates.

Benchmarking and evaluation cycles require data from the field, as well as human annotators, to ensure that the definition of "high quality" accurately reflects pedagogical best practices and students' real-life experiences. More data facilitates higher accuracy, but a major challenge is ensuring that data is cleaned and ready to use. Some organizations are sitting on substantial datasets that are not immediately useful, while others need to build infrastructure to make data interoperable or ready for analysis, properly stripped of sensitive information. Additionally, district and state privacy policies vary, making it difficult to scale data practices without robust data agreements. One developer described needing to "build a train out of the coal mine so data can be handled in a responsible way."

One solution to this challenge is using synthetic data in model evaluations. Synthetic data is AI-generated, which may not be as accurate as field data but can be scaled quickly with little cost. Inaccuracies can also be caught during evaluations with human graders, especially since many are expert educators who are intimately familiar with seeing student responses. Using this method, one organization has built reliable test cases of "thousands of different responses," which it uses to "regression test the newest AI models, updates, and prompting to make sure that feedback quality is really high."

Measuring AI performance is not always straightforward. For example, some organizations have found that giving too much context or information in the prompt can downgrade response quality rather than improve it as expected. Quality testing can also easily be mistaken for impact evaluations, making it sometimes unclear whether an ed tech tool is simply crafting outputs well versus truly making a difference for students. But measuring AI performance remains essential. Without rigorous testing, organizations risk deploying tools that produce inconsistent, inappropriate, or pedagogically unsound outputs — undermining the very learning goals they aim to support.

High-quality AI implementation also requires knowing when to hold back.

Just as important as where organizations use AI is where they deliberately choose *not* to use it. Across interviews, several clear patterns emerged around the boundaries developers set for AI use, given its current capabilities. In high-stakes, "deterministic" situations, most organizations refuse to use AI. This includes teacher evaluations used for employment decisions, and other contexts where errors or inconsistency could have serious consequences such as loss of funding.

Another clear boundary for multiple organizations is student-facing interactions. Most organizations, as described earlier, position teachers as the primary user and driver of learning. The consensus among these organizations is that AI is less useful — and for some, negative — for teaching children. One interviewee bluntly said, "I don't think a chatbot will set kids free. It won't be the most impactful, and we don't need to add more tech to their lives." A school leader also expressed disappointment with the landscape of student-facing tools because they only catered to a certain student: "… if a student has low skill and low motivation, tech doesn't serve students well. It's only serving students with high skills and high motivation well."

Overall, the principle guiding these choices among several organizations seems to be that AI is most valuable when it extends or amplifies what good teachers do, rather than attempting to replace or bypass their expertise. If the AI cannot enhance teaching practice or solve an actual problem better than existing solutions, these developers opt not to use it.

Sustaining Thoughtful Ed Tech in the Age of AI

As a somewhat nascent niche of the ed tech sector, AI-powered tools do not yet have a standard business model for long-term sustainability. One fundamental consideration is what is reasonable to charge school and district leaders with constrained budgets. Often, the "appropriate" price point varies by use case, guided by counterfactuals — in other words, if the administrators were not paying for the AI tool, what would be filling that gap? For classroom tools that augment teacher capabilities, administrators must weigh costs against time savings or improved personalization, factoring in number of students, session times, or grade levels. But for tools that address shortages — for example, school counselors or academic advisers — 15 minutes with a high-quality tool trained for academic or career advising is much easier to pay for than 15 minutes with a full- or part-time human counselor.

Among the factors complicating demand is support for implementation. Some organizations provide extensive support, seeing it as crucial to their theory of change, while others are trying to reduce dependence on intensive support to reach sustainability. Several are partnering with larger organizations such as curriculum providers or workforce training programs to expand their reach without incurring extra costs.

Looking forward, several developers noted substantial demand for real-time, live support for teachers, which may represent both an opportunity and a challenge for operational planning (Sidebar). At the same time, many interviewees described needing to balance requests for certain features with core theories of impact. While some feature requests sound exciting, often the fundamental question remains: "Is this something that teachers will actually use, or is this just something that sounds cool?"

On the supply side, a clear trend is that organizations are experimenting substantially and absorbing costs while trying to find product-market fit. As one leader described it, "Year 1 is seeing what sticks in learning before scaling for cost. We're going output by output and solution by solution, dabbling with paid versus free and different models." Another leader referenced a business framework that prioritizes customer satisfaction and demand over efficiency, which leaves costs as the least important consideration during the development phase. **As AI tool adoption continues to expand and organizations mature,[9] striking the right balance in managing technological costs to scale high-quality tools will be critical to ensuring that AI amplifies learning instead of diminishing it.**

SIDEBAR

What might the evolving AI-powered ed tech market look like in the future?

The ed tech landscape already includes several established technology providers, from school information systems (e.g., PowerSchool) to learning management systems (e.g., Canvas) and major publishing companies (e.g., HMH). Meanwhile, tech titans like OpenAI, Anthropic, Microsoft, and Google are increasingly turning their attention to education with tools such as ChatGPT's Study Mode, Claude's Learn Mode, Khanmigo, and Gemini in the Classroom.[10] Both the education incumbents and the AI power players command a wealth of resources across engineering talent, capital, and market share. As a result, smaller organizations may easily be pushed out or absorbed; as one developer put it, "We're all rounding errors to Google. They could crush any ed tech company." A natural question then arises: What will the future ed tech landscape look like as GenAI keeps evolving?

One scenario is akin to the status quo, where all three types of ed tech providers — incumbents, tech giants, and startups — continue to coexist, each offering distinct value propositions. A second scenario might see smaller ed tech providers acquired either by the education incumbents or by the larger tech companies. Absorption into the existing ed tech providers could accelerate adoption given established distribution channels but also dilute innovation, as specialized tools have to fit within existing systems. Absorption into the tech companies could also accelerate adoption, but perhaps unevenly without the deep pedagogical expertise from smaller founders. And a third scenario could see tech giants dominate the market, perhaps due to integrating education-related AI into their widely used platforms, as Google is starting to do with Gemini in the Classroom. Alternatively, advancements in frontier GenAI models might render education-specific tools moot — the general-purpose AI models could do it all.

There are many challenges looming before the ed tech sector, including financial turmoil and tool overload for teachers and districts as the market becomes increasingly crowded. These pressures may accelerate consolidation, but they also create opportunities for differentiation. Some interviewees suggested that their deep pedagogical expertise gives their tools an edge: One leader asserted that "there's something taste-wise that comes from deeply investigating a particular problem … Google would have to dive super deep into this particular problem space [to be compelling]." Other leaders see close relationships with schools and districts as another competitive advantage. As the market evolves, maintaining focus on the thoughtful, intentional design principles outlined in this series — regardless of which organizations ultimately deliver them — will be critical for ensuring AI genuinely improves teaching and learning.

Conclusion

Certain patterns and consensus are emerging among organizations incorporating AI in their ed tech. For example, there is widespread recognition that technology should support, not replace, teacher judgment and expertise. There is general agreement that productive struggle is central to learning and that AI tools must be designed to preserve appropriate challenge rather than eliminate it. While approaches to curriculum vary, organizations are thinking carefully about how AI tools interact with existing instructional materials rather than assuming AI should generate everything from scratch. And developers are grappling with measuring impact on students, even as they navigate significant resource constraints and the rapid pace of technological change.

The technical and operational considerations of implementing AI are numerous. Developers must balance questions of cost, privacy, accessibility, and quality while trying to build sustainable business models in a resource-constrained market. They are constantly testing AI performance quality and iterating rapidly as models and capabilities evolve. And perhaps most significantly, they are also setting boundaries around AI use — choosing not to apply it in high-stakes contexts or where it cannot genuinely replicate good teaching. This thoughtfulness, combined with deep educational expertise and relationships with schools and districts, may be an important differentiator as larger technology companies enter the education market.

Yet this is still early days: AI capabilities are evolving rapidly, long-term impact data remains limited, and market dynamics continue to shift. The organizations interviewed are navigating these complex trade-offs without clear road maps, but their thoughtfulness about pedagogy, measurement, and boundaries offers important lessons. As AI in education matures, the ongoing challenge will be sustaining this focus on high quality and measurable impact to ensure AI becomes a genuine lever for improving learning. ✦

Endnotes

1 Amy Chen Kulesa, Marisa Mission, Michelle Croft, and Mary K. Wells, *Productive Struggle: How Artificial Intelligence Is Changing Learning, Effort, and Youth Development in Education* (Bellwether, June 2025), https://bellwether.org/publications/productive-struggle/; Michelle Croft, Amy Chen Kulesa, Marisa Mission, and Mary K. Wells, *Measuring Artificial Intelligence in Education* (Bellwether, October 2025), https://bellwether.org/publications/measuring-ai-in-education/.

2 Croft, Chen Kulesa, Mission, and Wells, *Measuring Artificial Intelligence in Education*.

3 Tokens are units of text processed by LLMs and are the pricing metric used by most frontier LLM providers to access models directly through an API (rather than through a chatbot). For example, using the OpenAI API to send requests directly to GPT-5 costs $1.25 per 1 million tokens. "Cost Per Token," Tetrate, https://tetrate.io/learn/ai/cost-per-token; "API Pricing," OpenAI, https://openai.com/api/pricing/.

4 Amy Chen Kulesa et al., *Learning Systems: Shaping the Role of Artificial Intelligence in Education* (Bellwether, June 2025), 30, https://bellwether.org/publications/learning-systems/.

5 Ibid.

6 Davide Castelvecchi, "Can We Open the Black Box of AI?," *Nature News* 538, no. 7623 (October 2016): 20–23, https://www.nature.com/news/can-weopen-the-black-box-of-ai-1.20731; Vikas Hassija et al., "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognitive Computation* 16 (January 2024): 45–74, https://doi.org/10.1007/s12559-023-10179-8.

7 See for example: "AI Benchmarks for Education," AI-for-Education.org, https://ai-for-education.org/ai-benchmarks-for-education/.

8 Alex Spurrier and Marisa Mission, "The Leading Indicator: AI in Education Issue Thirteen," Bellwether, September 4, 2025, https://bellwether.org/ai-newsletter/the-leading-indicator-ai-in-education-issue-thirteen/.

9 Elizabeth Laird, Maddy Dwyer, and Hannah Quay-de la Vallee, "Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students," Center for Democracy & Technology, October 8, 2025, 6–10, https://cdt.org/insights/hand-in-hand-schools-embrace-of-ai-connected-to-increased-risks-to-students/.

10 "Introducing Study Mode," OpenAI, July 29, 2025, https://openai.com/index/chatgpt-study-mode/; "Introducing Claude for Education," Anthropic, April 2, 2025, https://www.anthropic.com/news/introducing-claude-for-education; "Gemini for Education," Google, https://edu.google.com/ai/gemini-for-education/.

About the Authors

MARISA MISSION

Marisa Mission is a senior analyst at Bellwether. She can be reached at **marisa.mission@bellwether.org**.

MICHELLE CROFT

Michelle Croft is an associate partner at Bellwether. She can be reached at **michelle.croft@bellwether.org**.

AMY CHEN KULESA

Amy Chen Kulesa is a senior associate partner at Bellwether and leads the organization's work on AI. She can be reached at **amy.chenkulesa@bellwether.org**.

About Bellwether

Bellwether is a national nonprofit that works to transform education to ensure young people — especially those furthest from opportunity — achieve outcomes that lead to fulfilling lives and flourishing communities. Founded in 2010, we help mission-driven partners accelerate their impact, inform and influence policy and program design, and bring leaders together to drive change on education's most pressing challenges. For more, visit **bellwether.org**.

ACKNOWLEDGMENTS

We would like to thank the many experts who gave their time and shared their knowledge with us to inform our work. Thank you also to the Bezos Family Foundation, Chan Zuckerberg Initiative, Charter School Growth Fund, and Overdeck Family Foundation for their financial support of this project.

We would also like to thank our Bellwether colleagues Mark Baxter and Mary K. Wells for their input and former colleague Janine Sandy for her support. Thank you to Amy Ribock, Kate Stein, Andy Jacob, McKenzie Maxson, Temim Fruchter, Julie Nguyen, and Amber Walker for shepherding and disseminating this work, and to Super Copy Editors.

The contributions of these individuals and entities significantly enhanced our work; however, any errors in fact or analysis remain the responsibility of the authors.

Bellwether

FORWARD THINKING. FORWARD MOVING.