Introduction

For years, math-focused, K-12 ed tech tools assessed students using multiple-choice questions — easy to grade at scale, but limited in gathering meaningful insights about learning and mastery. For students who struggle, this approach can reinforce deficit-based mindsets, as teachers see wrong answers without understanding where misconceptions originated or what partial knowledge students possess. When generative artificial intelligence (GenAI) arrived in 2022, EdLight Founder Teryn Thomas saw an opportunity to change this status quo. Instead of multiple choice questions, EdLight uses GenAI to analyze and provide feedback on students' handwritten math work, helping teachers understand student thinking at scale. In Thomas' words, EdLight helps educators seek "the area between right and wrong," giving teachers more agency than they might otherwise feel with traditional structures and data.[1]

This focus on nuanced understanding reflects a core belief that teachers with greater capacity and more intuitive ways of interpreting data will feel more empowered to influence students' outcomes. EdLight uses GenAI to "see" a student's handwritten work and instantly diagnose the student's strategy, whether they were successful, and — if they weren't — which misconceptions might have prevented the student from succeeding. From there, the tool creates "tailored, tangible, and actionable insights" for both students and teachers.[2]

While feedback is a well-tested method for boosting student learning,[3] and many ed tech developers are using GenAI models to create feedback for students, EdLight's approach to using AI stands out for its:

1. Tight logic model that centers educators, guiding design and product decisions to preserve teachers' judgment.
2. Purpose-built training dataset that centers students furthest from opportunity to ensure accessibility for all.
3. Self-hosted and self-trained models that protect student data and ensure high-quality, tailored responses.
4. Grounding in deep expertise in math reasoning and pedagogy.

From Issue to Impact

EdLight works to increase a teacher's capacity to serve both individual students and the whole classroom.

In the classroom, teachers often lack the time to a) provide timely, individualized feedback on open-ended, handwritten work, and b) synthesize patterns across a class to judge overall mastery and common misconceptions. At the district level, teacher shortages — especially in critical areas such as special education or science, technology, engineering, and math (STEM) — create large classrooms and high caseloads that further limit teacher capacity.[4]

To minimize educators' tradeoffs, EdLight instantly analyzes handwritten work to provide insights at both the student and classroom level. For each student's submission, the system produces a summary of the student's work, a list of misconceptions that led to errors (when applicable) and next steps the teacher might take, such as asking additional questions or providing different examples.[5] Dashboards then aggregate the student-level insights to reveal common misunderstandings across the class.[6] The classroom-level view also generates a summary that allows teachers to more precisely understand the collective mastery level: For example, rather than providing a message like "70% of students did not complete the problem correctly," EdLight might offer a message like "60% of students were very close and only had computational errors, but two students need significant intervention."[7] By automating feedback generation and progress tracking in a way that creates actionable insights and recommendations, the tool saves teachers between 2 and 5 minutes per assignment.[8]

Throughout development, EdLight's designers kept student accessibility in mind.

The first step to using EdLight is scanning a student's handwritten work, which can be done with a Chromebook or tablet by either a student or teacher. An AI model then uses optical character recognition (OCR) to understand what's on the page — a surprisingly complex endeavor. OCR has existed for decades,[9] but it can struggle to recognize less-than-perfect handwriting.[10] EdLight's focus on math added another layer of complexity: The OCR needed to be able to recognize the symbols, figures, and unstructured work inherent to math — not just linear text. And critically, the OCR needed to be able to recognize handwriting from students of all different backgrounds without changing how their work was scanned.

To address these challenges, Thomas adopted a universal design mindset, which encourages creating products that can be used by the widest possible range of people without requiring adaptation. EdLight used a $3 million grant to collect more than 300,000 samples of handwritten student work, which formed a repository of data that could train the model's OCR capabilities. The EdLight team intentionally curated this dataset to center students from underserved backgrounds (e.g., students from low-income households, students with individualized education programs (IEPs), students with functional learning disabilities, students who struggle with executive functioning, multilingual students). Each student sample was then annotated by hand, by human educators. This process, although time- and resource-intensive, created a highly-accurate training dataset that made EdLight's AI model able to parse nearly any student's handwriting. By focusing on students furthest from opportunity, EdLight's designers created a tool more accessible for everyone.

EdLight continues to maintain and grow the repository through data-sharing agreements with districts and schools in Washington, D.C. and 13 states across the country.[11] This expansion of student work will allow the tool to keep improving its OCR abilities and remain accessible to all students and teachers, regardless of background.

EdLight identifies educators as a linchpin for student success and long-term impact.

Rather than replacing teachers, EdLight relies on educators as the main vehicle for changing student outcomes.[12] The tool's goal is to alleviate both the immediate teacher capacity crunch and second-order effects such as frustration, lack of motivation, and burnout.[13] Over time, the tool's assistance for teachers will increase their effectiveness and thereby improve students' academic progress and outcomes.[14]

This theory is clearly illustrated in EdLight's logic model, which identifies immediate outputs (e.g., usage rates, feedback delivery success, formative data) as well as short-term, medium-term, and long-term outcomes at both the student and teacher levels (Table). The model draws clear connections between EdLight usage and expected changes, which helps the organization a) articulate how their use of GenAI is meaningful, and b) measure whether the product is performing as intended.

Table. Example measures of success from EdLight's logic model[15]

	Immediate Outputs	Short-Term Outcomes	Medium-Term Outcomes	Long-Term Outcomes
Student	• Assignment completion • Misconceptions identified • Formative assessments	• Increased math self-efficacy • Increased motivation • Increased persistence through additional revisions or attempts	• Increased student proficiency in math, as measured by state assessments • Additional opportunities for students to complete on-grade level work	• Improved student academic outcomes and graduation rates • Shrinking disparities in outcomes across student demographics
Educators	• Usage rates of EdLight for grading • Uptake of AI-generated recommendations or insights	• Increased teacher confidence in using formative assessment data • Increased time to use on instructional preparation or professional development	• Improved teacher effectiveness • Greater visibility into trends across classes, teachers, and grade levels	• Decreased educator burnout • Improved teacher retention rates

Some schools that are using EdLight have already seen shifts in teacher behavior aligned to EdLight's expected short-term outcomes. One partner school reported that over one school year, EdLight helped 94% of teachers increase their use of formative assessments, and across the school, leaders saw "a significant shift in their professional collaboration and engagement with data."[16] Another school estimated that its teachers saved "between 3,000 and 7,500 hours [of grading] over two years" with EdLight, allowing them to "invest more deeply in instructional planning."[17] While additional evaluations are necessary to understand EdLight's impact over time, these early case studies are promising initial evidence of the tool's efficacy.

Using AI With Intention

EdLight's backend infrastructure showcases intentional design choices guided by high expectations for quality.

Instead of a "single shot" approach where one large language model (LLM) is carefully customized for the desired output, or a multi-model approach that uses out-of-the-box LLMs (e.g., ChatGPT, Claude, Gemini), EdLight is powered by a string of specialized GenAI models that each play a different role. This "pipeline" of models is advantageous for a few reasons: First, given the speed at which GenAI is advancing, breaking the process into parts executed by different models allows EdLight to update parts of the pipeline as relevant GenAI advances are released, without being vulnerable to model rollbacks or crashes, reliant on updates from major developers, or subject to take the entire system offline for updates. Second, the pipeline approach gives EdLight greater control over the quality of the output: Instead of sending student work into a "black box," EdLight can trace back along the pipeline to determine where an error might have led to a faulty output. Lastly, using multiple models allows EdLight to tap specialized versions for higher-quality outputs. For example, the first model that scans a student's worksheet specializes in OCR and is trained on the dataset described earlier. It is separate from the model that analyzes the student's work for misconceptions, which is trained specifically for mathematical reasoning.

Most of EdLight's pipeline is powered by different versions of Qwen, one of the only multimodal models that is open-source. Unlike proprietary frontier models, open-source models can be downloaded, trained, and hosted on custom servers, which allows for more customization and control. For EdLight, self-hosting their models also provides better privacy protection. Because all of the computing happens on EdLight servers, the organization can set boundaries to control how data is processed, what is kept, and who can access that data. This infrastructure is much more complex than most ed tech tools' backends; however, everything about it — from the pipeline of specialized models to the custom servers — demonstrates EdLight's focus on quality, student and teacher privacy, and safety.

EdLight fine-tuned their own models to ensure quality, rather than relying on well-known frontier models.

Fine-tuning is the process of adapting a generalized AI model to a specific task or context by training it on a carefully chosen set of relevant examples.[18] For example, the out-of-the-box version of the model EdLight uses to scan student work already had OCR capabilities and could likely recognize most handwriting samples. However, because EdLight's use case was more specific (math problems in handwriting from students of all backgrounds), the model had to be fine-tuned using the annotated dataset of handwritten student math work. That training taught EdLight's model to recognize a wider range of handwriting styles, especially from younger students and students from underserved backgrounds.

EdLight also fine-tunes the other models in the pipeline using teacher feedback, internal benchmarks, and specific K-12 math problems. Thomas sees the process as essential to ensuring that the models are using the best practices in math pedagogy and feedback. In her words, the AI models must be capable of replicating a teacher's ability to "look at a student's worksheet, understand their chain of thought, diagnose any misconceptions, and respond accordingly — otherwise, the technology is not good enough to make a difference."[19]

EdLight's focus on quality leads to tradeoffs in costs.

Self-hosting and fine-tuning models are rare practices, especially for ed tech developers, because both come with high price tags. To be able to host, train, and run all of their models, EdLight invested in their own graphics processing units (GPUs) as well as set up their own servers — infrastructure with very high fixed costs. Fine-tuning EdLight's models also required specialized machine learning expertise, especially when it came to training the OCR model for K-12 math. The most similar GenAI tool approaches could only be found in health care or higher education; as Thomas noted, "Not a ton of folks are obsessed with solving this for a second grader."[20] Given the high demand for AI and machine learning talent, EdLight also faced significant startup costs there.

Having specialized models can also make it difficult to improve them without sacrificing latency (speed). For example, when EdLight updated some of their models to see a 10-percentage-point increase in accuracy, the change added an additional 90 seconds of processing time — not ideal for teachers who may need to respond in real time in the classroom. Improving latency, however, requires additional computing capacity, which means additional GPUs and additional costs. Thomas recounted the challenge, saying "It feels like an easy decision [to add more GPUs] but now that has a real financial implication."[21]

Due to the high fixed costs and added complexity, self-hosting and fine-tuning are uncommon for ed tech developers. Those creating applications for text-based subjects (e.g., reading, history) might not see either practice as necessary given that LLMs are naturally strong in text-based inputs and outputs. Others may struggle to front the initial costs. But the team at EdLight intentionally prioritized student privacy, quality control, and integration of pedagogical best practices, and they were able to leverage venture capital and philanthropic funding to meet those goals.[22]

Amplifying Learning

Technology-driven automation often brings fears of diluting a student or teacher's "productive struggle," or the effort required to engage with challenging problems that ultimately boosts learning more than if the struggle were not required. However, EdLight's approach to product design showcases a few ways developers can use AI to *amplify* productive struggle and therefore learning:

1. **Feedback mechanisms encourage students to do the thinking first.** The core premise of EdLight is for students to attempt the work first, before they can receive feedback. Students therefore must use their working memory or actively recall prior knowledge to solve the work, which promotes more enduring encoding of information into their long-term memory.
2. **Thoughtful feedback calibrates the amount of struggle a student faces at any given time.** The most learning happens when students face a challenge just beyond their reach; too much struggle can lead to overwhelm and shutdown.[23] Because EdLight's math reasoning models identify root misconceptions, the resulting feedback is targeted to where a student is struggling the most, minimizing the risk for overwhelm.
3. **Continuous practice cultivates motivation and a growth mindset.** When EdLight is used during class or homework as practice, students can redo problems and resubmit their answers until they master the concept. This process builds their growth mindset and generates intrinsic motivation as students experience their own progress.
4. **Naming misconceptions can encourage healthy metacognitive practices.** As students receive feedback tailored to their specific misconceptions, they are building metacognitive skills, including

learning to recognize what they did wrong, why it was wrong, and how they might approach a similar problem in the future.

Whether an AI tool amplifies productive struggle can be a key indicator of its ability to drive student learning and ultimately, student outcomes. EdLight is starting to reap the benefits of its design: Brown University researchers recently found that the insights generated by EdLight's analyses of student work strongly correlated with end-of-year academic outcomes, with predictive validity — indicating that EdLight data is a valuable tool for gauging student mastery and progress.[24]

Conclusion

EdLight is continuing to build features for both teachers and students that will capitalize on the data its GenAI models generate. On the horizon are tools for teacher professional development (including an AI-powered instructional coach); a pilot of a live dashboard that updates as students complete their work in class; and a partnership with the Math Narrative Project to further fine-tune EdLight's models so that the feedback generated incorporates ways to boost students' math-related self-efficacy.

Designing an ed tech tool that uses GenAI will always come with tradeoffs among speed, quality, sustainability, privacy, accessibility, and more. EdLight's robust backend infrastructure illustrates a set of design choices that center widespread accessibility, high-quality feedback, teacher growth, and data privacy — despite the resulting higher costs. These choices were intentional and may be instructive both for ed tech developers using similar principles and for those hoping to better understand how GenAI might truly promote learning rather than just introducing novelty.

Acknowledgments

We would like to thank the experts who gave their time and shared their knowledge with us to inform our work, including Teryn Thomas and EdLight staff. 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 for his input and [PRIMARY EA] for her/his/their support. Thank you to Amy Ribock, Kate Neifeld, Andy Jacob, McKenzie Maxson, 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.

About the Authors

In preferred author order, insert names and short bios below.
Find Headshots and Bios in the Brand Identity Toolkit (headshots are not needed).

MARISA MISSION
Marisa Mission is a senior analyst at Bellwether in the Policy and Evaluation practice area. She can be reached at marisa.mission@bellwether.org.

MICHELLE CROFT
Michelle Croft is an associate partner at Bellwether in the Policy and Evaluation practice area. She can be reached at michelle.croft@bellwether.org.

AMY CHEN KULESA
Amy Chen Kulesa is a senior associate partner at Bellwether in the Strategic Advising practice area. She can be reached at amy.chenkulesa@bellwether.org.

About Bellwether

Bellwether is a national nonprofit that exists to transform education to ensure systemically marginalized young people achieve outcomes that lead to fulfilling lives and flourishing communities. Founded in 2010, we work hand in hand with education leaders and organizations to accelerate their impact, inform and influence policy and program design, and share what we learn along the way. For more, visit bellwether.org.

Creative Commons