EdLight Research Portal: An Expert-Annotated Repository of Handwritten Math Student Work

Meneses, M. & Castro, S.

formative assessments | Handwritten math dataset | K-12 student reasoning

This paper introduces the EdLight Research Portal (ERP), a publicly available repository of expert annotated handwritten math student work. ERP addresses the gap between formative assessment practices in real mathematics classrooms and the digital datasets commonly used in AI-powered education research. The dataset captures authentic paper-based reasoning from 5th to 9th grade students and includes standardized rubric scores, misconception tags, and contextual metadata generated by trained math educators. ERP enables researchers to study student thinking at scale, examine patterns in mathematical misconceptions, and benchmark machine learning models on tasks that reflect real classroom data. The portal provides a foundation for advancing transparent, pedagogically grounded research in AI applications for mathematics education.

Introduction

Learning Engineering (LE) integrates learning sciences with engineering rigor to iteratively improve educational systems through data-informed practice (Baker et al., 2022). In K-12 mathematics, LE can alleviate teacher burdens, such as limited instructional time, by automating routine assessments and refining recommendation systems (Kadaruddin, 2023; Noroozi et al., 2024). However, rapid LE research is currently hindered by a lack of high-quality, annotated datasets that reflect authentic classroom conditions. While existing repositories often rely on typed or digital ink inputs (Gervais et al., 2024; Zhang et al., 2024), they fail to capture the nuanced reasoning found in traditional handwritten work, the primary mode of formative assessment in schools.

To bridge this gap, we introduce the EdLight Research Portal (ERP): a public repository of 19K expert-annotated samples of authentic handwritten K-12 mathematics. Developed as a nested LE cycle within the broader EdLight platform, each sample features rubric scores, misconception tags, and metadata produced by trained educators. By providing this domain-specific infrastructure, we enable researchers to study student thinking at scale, identify mathematical misconception patterns, and benchmark machine learning models against data that truly reflects the classroom experience.

Methods

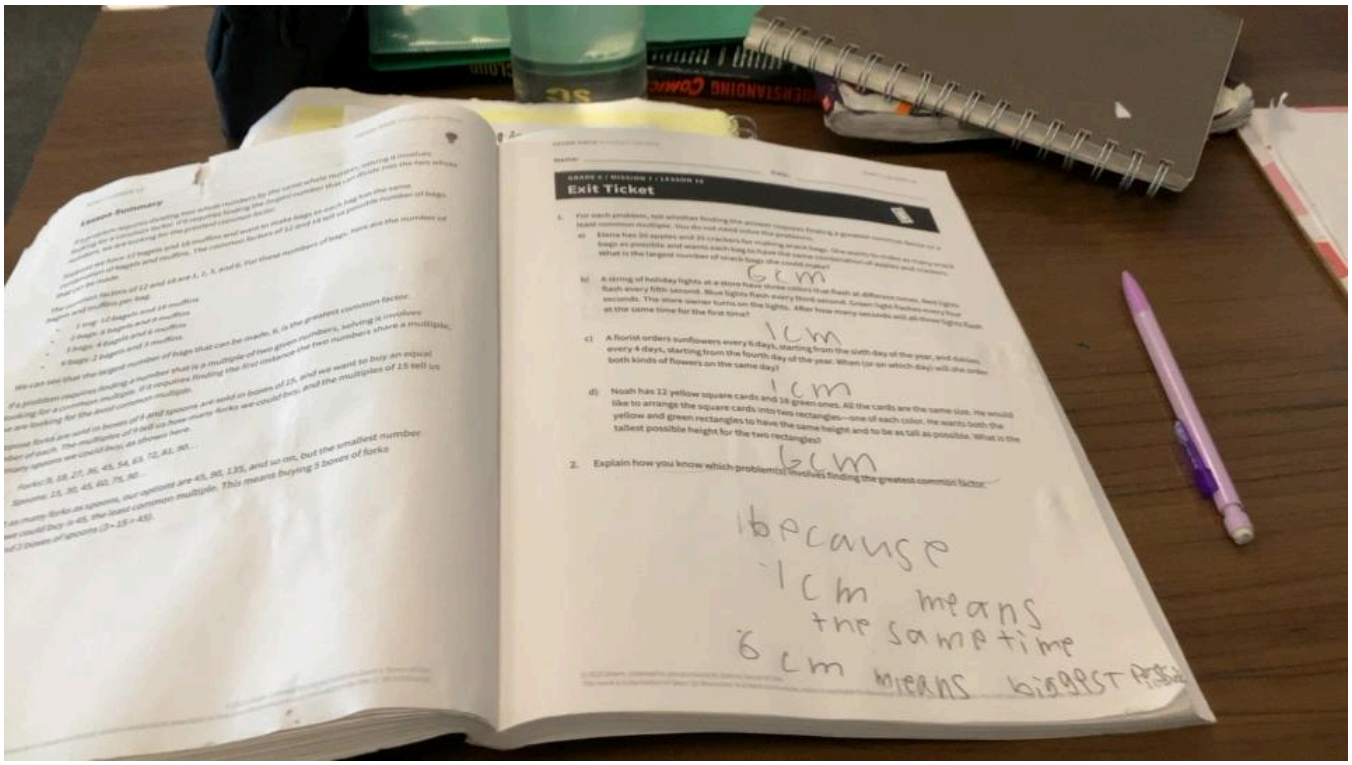
Data Collection and Preparation

The dataset was collected in collaboration with three EdLight partner schools, where students completed paper-and-pencil mathematics assignments during regular classroom instruction. These artifacts were scanned or photographed by teachers using EdLight's platform and securely uploaded to a centralized repository. Each submission was paired with metadata provided by teachers, including grade level, assignment details, learning standards, and anonymized student demographics when available. All student work samples as of November 2025 come from 5th to 9th grade Illustrative Mathematics exit tickets, known as Cool Downs.

To ensure high-quality annotations at scale, EdLight hired a team of 55 experienced K-12 math teachers, referred to as math instructional specialists (MIS), who conducted all labeling activities between November 2023 and June 2024. These specialists were trained to assign misconception tags and overall rubric scores based on standardized scoring guidelines aligned with state and national math standards. Their work produced expert-labeled samples that preserve the diversity and authenticity of real student reasoning while maintaining consistency across thousands of annotations.

Figure 1.

A student-work image sample from the released dataset. It corresponds to an Illustrative Mathematics exit ticket (i.e., "Cool Down") for Lesson 18 of Unit 7 in 6th grade.



Annotation of Student Work

Each student response was evaluated along two dimensions: overall rubric score and mathematical misconceptions. Rubric scores range from 0 to 3, where 3 indicates a complete and correct solution, 2 reflects conceptual understanding with minor errors or incomplete reasoning, 1 indicates significant errors, and 0 shows no viable mathematical strategy.

Misconception annotations were developed in collaboration with district math leaders and one of the authors of the Illustrative Mathematics middle school curriculum. This collaboration ensured that the schema aligns with instructional expertise, reflects curricular intent, and captures the recurring reasoning and procedural errors students commonly make. Responses could receive multiple tags, including repeated instances of the same category, enabling granular characterization of student thinking. Responses without observable errors were tagged "N/A – No misconceptions present." This dual-layer approach captures both correctness and reasoning quality, supporting analyses beyond binary right-or-wrong evaluation.

Table 1.
Misconception tags developed by EdLight.

Misconception	Definition
Computation error	Mistakes in arithmetic operations such as addition, subtraction, multiplication, or division.
Precision error	Mistakes related to numerical accuracy or formatting.
Conceptual misunderstanding	A lack of understanding or incorrect interpretation of mathematical concepts or principles.

Representational error	Mistakes in expressing or interpreting mathematical information through visual representations.
Insufficient explanation	Inadequate explanation or model.
Incomplete	Students didn't finish or didn't answer the question at all.
Did not follow	Refers to cases where students did not follow directions.
Unable to diagnose	There is insufficient visible evidence to determine any other type of misconception.

Quality Assurance and Scoring Reliability

A structured quality assurance process was implemented to ensure consistency in both overall rubric scoring and misconception tagging. During the initial phase of annotation, team members manually conducted inter-rater reliability checks to align interpretations of the scoring criteria and misconception definitions. This calibration process included independent scoring of shared samples and group discussions to resolve discrepancies and refine annotation guidelines.

Following the transition to a web-based data annotation platform, quality assurance procedures were maintained through systematic cross-annotation and expert auditing. A subset of student responses was assigned to multiple annotators to monitor agreement across scorers. Senior instructional specialists conducted weekly audits, reviewing approximately ten percent of all completed annotations to verify adherence to scoring standards and misconception criteria. Feedback from these audits was shared regularly with annotators and informed ongoing refinements to training materials and annotation procedures.

Results

The EdLight Research Portal (ERP) is publicly available at https://research-portal.edlight.com (access can be requested via the form at https://api.edlight.com/contact/). ERP contains 19,641 samples of handwritten mathematics student work within 5th to 9th grades. Each sample includes a link to the corresponding image, metadata such as assignment name, content standard, grade level, and available demographic information including student gender and ethnicity. In addition, all samples are annotated with expert-generated misconception tags and overall rubric scores. Table 2 summarizes the distribution of samples across grade levels and mathematical content areas.

Table 2.
Count and percentage of ERP's samples across content areas by grade.

Content area	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9
Expressions & equations	198 (30%)	17 (3%)	4,446 (33%)	1,486 (31%)	166 (87%)
Functions	-	-	-	158 (3%)	15 (8%)
Geometry	48	-	1,030	1,462	8

	(7%)		(8%)	(31%)	(4%)
Number System	79 (12%)	373 (66%)	5,014 (37%)	642 (13%)	2 (1%)
Ratios & Proportions	342 (51%)	173 (31%)	2,665 (20%)	726 (15%)	-
Statistics & Probabilities	-	-	284 (2%)	307 (6%)	-
Total student-assignments	667 (100%)	563 (100%)	13,439 (100%)	4,781 (100%)	191 (100%)

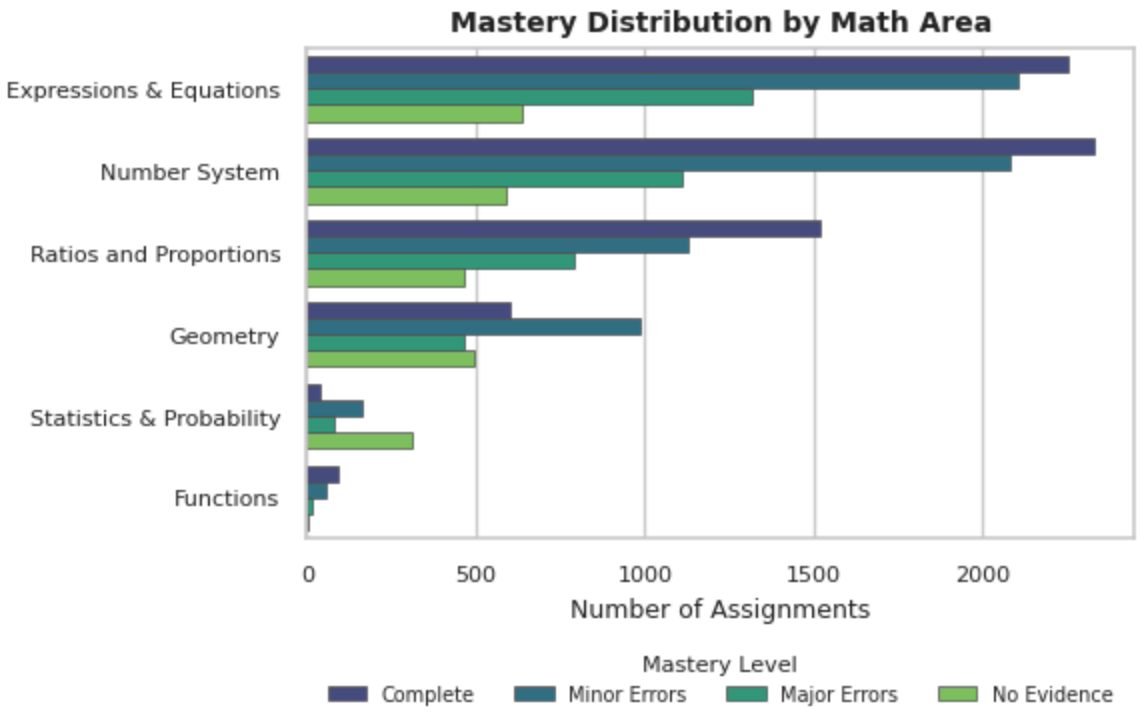
Table 3 presents the frequency and proportion of misconception tags assigned to student responses. Because a single sample may contain more than one type of misunderstanding, multiple tags can be applied to the same piece of work. Figure 2 complements this analysis by showing the distribution of overall rubric scores across content areas, allowing for comparisons between student performance and the prevalence of conceptual challenges.

Table 3.
Frequency and proportion of misconception tags assigned to student responses in the EdLight Research Portal. Multiple tags can be applied to a single sample, reflecting the presence of multiple reasoning errors within individual student work.

Misconception	Percent of student work
Computation error	11%
Precision error	4%
Conceptual misunderstanding	38%
Representational error	4%
Insufficient explanation	12%
Incomplete	6%
Did not answer specific question	2%
Unable to diagnose	11%
No misconception present	36%

Figure 2.

Distribution of overall rubric scores across content areas in the EdLight Research Portal, illustrating variations in student performance and enabling comparison with the prevalence of annotated misconceptions. Mastery level descriptors are abbreviated for legibility.



Discussion

The EdLight Research Portal provides foundational infrastructure for Learning Engineering research in K–12 mathematics by capturing authentic paper based student work at scale. By preserving handwritten reasoning together with rubric aligned evaluations, misconception tags, and classroom relevant metadata, ERP reflects formative assessment as it occurs in everyday instructional practice. This fidelity to classroom conditions addresses a central challenge in Learning Engineering by enabling data informed, iterative improvement of educational systems grounded in real student thinking.

As a public, educator annotated resource embedded within an ongoing Learning Engineering cycle, ERP supports the study of learning progressions, misconception patterns, and instructional decision making in authentic contexts. It also enables the development and evaluation of computational tools, including AI assisted assessment, that are designed to reduce teacher burden while remaining aligned with instructional intent. Future work will expand the dataset with additional solution level annotations and examine how such tools can be integrated into formative assessment workflows to support scalable improvement in mathematics education.

Acknowledgments

We thank the Gates Foundation for supporting this work, the Research Partnership for Professional Learning (RPPL) and Annenberg Institute at Brown University for analyses informing the dataset design, HackSoft for engineering support, and the educators and district partners whose expertise made this dataset possible.

References

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior, 3(1), 1–23. https://doi.org/10.1037/tmb0000058

Gervais, P., Fadeeva, A., & Maksai, A. (2024). MathWriting: A Dataset For Handwritten Mathematical Expression Recognition. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2.

Kadaruddin, K. (2023). Empowering Education through Generative AI: Innovative Instructional Strategies for Tomorrow's Learners. International Journal of Business, Law, and Education.

Noroozi, Omid et al. (2024). Generative AI in Education: Pedagogical, Theoretical, and Methodological Perspectives. International Journal of Technology in Education.

Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K., Gao, P., & Li, H. (2024). MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? European Conference on Computer Vision.