

Building Teacher Capacity to Generate and Evaluate Math Items with Small LLMs

Minseok Kim

University of Pennsylvania
mk8393@upenn.edu

Abstract

Automatic item generation with large language models typically depends on proprietary, cloud-hosted systems, which raises cost, privacy, and equity concerns for schools. Small language models (SLMs) run locally and reduce these concerns, but they also produce mathematical and instructional errors that require teacher oversight. This study proposes a measurement-grounded professional development (PD) model in which upper-elementary teachers use local SLMs to generate and evaluate mathematics items. As a foundation for the planned teacher study, we coded 75 Grade 3–5 items generated by three SLMs using a six-dimension teacher-facing rubric. Only 37 of the 75 items had a correct designated answer, only 8 of those 37 also had three diagnostic distractors, and only 7 of the 75 items passed all six dimensions. These findings motivate a productive imperfection approach, in which model failures become structured opportunities to develop teachers' AI judgment and assessment literacy.

1 Introduction

Automatic item generation (AIG) develops test items by specifying their content and intended cognitive demand (Gierl & Haladyna, 2012). Large language models (LLMs) have expanded these capabilities, but most demonstrations rely on large proprietary, cloud-hosted models. For schools, this dependence raises recurring API costs, privacy and data-governance concerns under regulations such as FERPA, and equity risks for under-resourced districts.

Two gaps remain. First, existing work centers on system architecture and content-generation quality rather than how teachers develop the capacity to use and critically evaluate these tools. Reza et al. (2025), for instance, validate a teacher-facing framework's outputs but leave teacher learning and pedagogical impact to future work, a gap made

pressing because smaller models struggle more as tasks grow complex (Wang et al., 2025). Second, whereas traditional AIG applies multistage validity and psychometric scrutiny (Embretson & Kingston, 2018), evaluations of LLM-generated educational content often stop at surface-overlap metrics such as BLEU and ROUGE (e.g., Masum et al., 2025), without evidence of item validity, difficulty, or fairness. Teachers increasingly need both AI fluency and assessment literacy. The latter, Popham (2009) argues, must be supplied through in-service professional development (PD) because preservice programs rarely provide it.

This work has three aims. The first is to design a PD model in which elementary mathematics teachers use small, locally deployable LLMs to generate and evaluate instructional items. The second is to examine the quality and recurring failure patterns of SLM-generated items using a teacher-facing, measurement-grounded rubric. The third is to use these findings to inform a subsequent study of teacher AI fluency, assessment judgment, and item-evaluation expertise. At this work-in-progress stage, the study analyzes a controlled pilot corpus of 75 Grade 3–5 mathematics items generated by three small open-weight models across five task types. Teacher implementation is pending IRB approval. A subsequent phase will recruit in-service teachers to evaluate, revise, and provide feedback on these items.

2 Background

2.1 Automatic Item Generation and Small Language Models

Traditional AIG emphasizes explicit specifications for content, cognitive demand, and item quality. Generated middle-school mathematics items can meet the same review and psychometric standards as operational items (Embretson & Kingston, 2018).

Small language models (SLMs) offer a viable alternative. In education, locally deployable SLMs can perform structured analytic tasks, though they remain weaker at higher-level evaluative judgment (Favero et al., 2025), and run offline on consumer hardware via fine-tuning and quantization (Masum et al., 2025). Most directly, Reza et al. (2025) propose a teacher-centric framework in which local SLMs generate materials through an interactive refinement loop, framed as ensuring educator agency rather than a workaround for model limits.

2.2 Professional Development and Assessment Literacy

The PD is a multi-session model incorporating the core features of effective PD: content focus, active learning, coherence, duration, and collective participation in teacher learning communities (Desimone, 2009; Popham, 2009). Following Desimone’s (2009) causal-path model, we focus on the first links, from PD to changes in teacher knowledge, skills, and beliefs, not distal student outcomes (see Figure 1).

2.3 Productive Imperfection

A central design principle, **productive imperfection**, reframes model error as material for learning (Figure 2). This builds on evidence that errors aid learning mainly when learners possess sufficient prior knowledge to explain why an answer is wrong (Große & Renkl, 2007). Rather than treating these errors solely as defects to be filtered out, the PD uses them as contrasting cases that teachers must diagnose, explain, and revise. Teachers are particularly well positioned for this activity because their domain knowledge enables them to interrogate fluent but flawed outputs rather than accept them at face value. Productive imperfection therefore links model failure to deliberate practice in assessment judgment.

3 Methods

3.1 Professional Development Design

In early sessions, teachers will use a small open-weight model (e.g., Qwen3 8B, Gemma 3 4B, or Phi-4-mini; Qwen Team, 2025; Gemma Team, 2025; Microsoft, 2025) with structured prompts and item templates specifying grade, target standard, cognitive demand, and answer-key requirements. Prompt design and task decomposition strongly affect small-model performance (Favero et

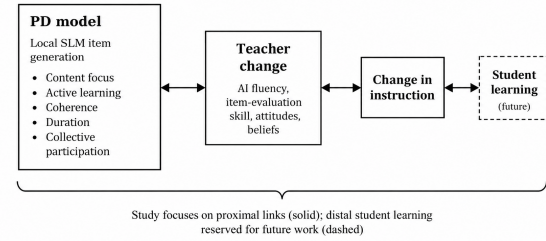


Figure 1: Study logic model adapted from Desimone (2009). Solid links represent the planned PD outcomes; the dashed student-learning link is reserved for future work.

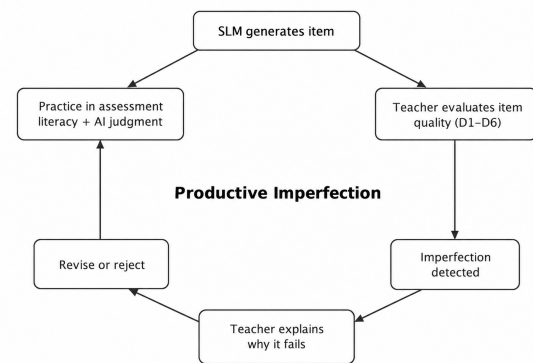


Figure 2: Productive Imperfection cycle for evaluating SLM-generated mathematics items.

al., 2025). In later sessions, teachers will apply an evaluation rubric grounded in assessment-literacy competencies (Popham, 2009) to judge item validity, accuracy, alignment, clarity, cognitive demand, and distractor quality, then revise or reject items.

The pilot items described below are intended to serve as authentic practice material, particularly because the models generated flawed items at a substantial rate on multi-step tasks (Tables 2 and 3). Teacher implementation is pending IRB approval; the present work-in-progress therefore establishes the technical, measurement, and design foundation for the subsequent PD study. The planned evaluation will examine teacher self-efficacy and AI comfort, item-evaluation accuracy on calibrated items drawn from the pilot corpus, and the quality of teacher-vetted item pools.

3.2 Pilot Generation Protocol

To establish the item pool for the planned teacher study, we generated a controlled pilot corpus using three small language models. All models were run locally in LM Studio with a 4,096-token context window. Qwen3 8B was run with thinking

disabled. Gemma 3 4B and Phi-4-mini were used in their standard instruction-tuned configurations. The three models are referred to below as Qwen, Gemma, and Phi.

Each model completed five sessions. In each session, the model generated one item for each of five task types: rounding, single-digit multiplication, fraction comparison, decimal place value, and multi-step fraction addition. The same prompts were used across models and sessions, yielding 75 Grade 3–5 items (3 models $\times$ 5 task types $\times$ 5 sessions).

The five prompts were issued sequentially within a single conversation. Later items were therefore conditioned on earlier turns and should not be treated as independent samples. An initial Gemma run used a 512-token context window. All 25 items were subsequently regenerated at 4,096 tokens, and only the regenerated set was included in the primary analysis.

3.3 Coding and Analysis

A single researcher coded all items in randomized, model-masked order and proceeded one dimension at a time across the corpus. Items were coded using a researcher-developed six-dimension rubric informed by assessment literacy, AIG, and cognitive-demand research (Popham, 2009; Embretson & Kingston, 2018; Anderson & Krathwohl, 2001): key correctness (D1), validity (D2), alignment (D3), demand (D4), distractor diagnosticity (D5), and clarity (D6). D1 was established before the final codebook freeze, while D2–D6 were coded afterward. Non-viable items were coded NA on D4 and excluded from that denominator. Operational decision rules for D1–D6 are provided in Appendix B.

Table 1: Researcher-developed teacher-facing item evaluation rubric.

Dimension	Criterion
D1 Key	Is the keyed answer correct?
D2 Validity	Is there exactly one correct, distinct, and unambiguously keyed option?
D3 Alignment	Does the item test the intended math skill?
D4 Demand	Does it require the intended level of thinking?
D5 Diagnosticity	Do wrong options reflect plausible student errors with sound supporting rationales?
D6 Clarity	Is the item clear and grade-appropriate?

Distractor diagnosticity (D5) was evaluated at

both the distractor and item levels. For distractor j in item i ,

$$d_{ij} = G_{1,ij}G_{2,ij}G_{3,ij}, \quad (1)$$

where $G_{1,ij}$, $G_{2,ij}$, and $G_{3,ij}$ are binary indicators of whether the distractor (G1) is reachable through an identifiable student error, (G2) matches the stated misconception, and (G3) is supported by a mathematically sound rationale, respectively. Item-level D5 was defined as

$$D5_i = \prod_{j=1}^3 d_{ij}, \quad (2)$$

such that an item passed D5 only when all three designated distractors were diagnostic. Non-viable items were coded NA on D5 and excluded from the denominator.

Model, session, and turn identifiers were re-joined after coding using stored item IDs, and mathematical calculations were verified by script. Judgment calls were documented in an ambiguity log without revising the frozen codebook.

As a descriptive supplement to D6, Flesch–Kincaid grade level was computed from item stems and options. Because mathematical notation can distort readability formulas, FK scores were not used as a pass/fail criterion.

4 Results

4.1 Overall Item Quality

Across the 75 generated items, 37 (49%) carried a mathematically correct designated answer. Answer-key accuracy alone, however, overstated diagnostic item quality: of the 37 key-correct items, only 8 (22%) also met the item-level distractor criterion, and none of the 34 evaluable items with an incorrect key did so. At the distractor level, 96 of 213 evaluable distractors (45%) were coded diagnostic. Only 7 of 75 items (9%) passed all six dimensions. Four items, all from Phi, were non-viable and were coded NA on cognitive demand and distractor diagnosticity; denominators are adjusted accordingly throughout.

An initial Gemma run used a 512-token context window, under which 21 of 25 generations exceeded the loaded context. After regenerating all 25 items at 4,096 tokens, answer-key accuracy remained low (12/25 vs. 9/25), suggesting that context length alone did not explain the failure pattern.

The 512-token set was used only as a sensitivity check.

Table 2: Model by dimension (D1–D6): items passing / items evaluable.

Model	D1	D2	D3	D4	D5	D6
Qwen3 8B	15/25	23/25	21/25	19/25	4/25	20/25
Gemma 3 4B	9/25	14/25	22/25	19/25	2/25	22/25
Phi-4-mini	13/25	18/25	18/25	17/21	2/21	14/25
Total	37/75	55/75	61/75	55/71	8/71	56/75

Note. Four non-viable Phi items were excluded from D4 and D5.

Table 3: Task type by dimension (D1–D6): items passing / items evaluable.

Task Type	D1	D2	D3	D4	D5	D6
Rounding	8/15	15/15	14/15	14/15	0/15	14/15
Multiplication	10/15	13/15	12/15	13/14	0/14	14/15
Fraction comp.	9/15	11/15	15/15	15/15	3/15	10/15
Place value	9/15	11/15	12/15	9/14	5/14	10/15
Multi-step frac.	1/15	5/15	8/15	4/13	0/13	8/15

Note. Reduced D4 and D5 denominators reflect non-viable items.

4.2 Model and Task Patterns

Performance varied by dimension rather than producing a single ranking (Table 2). Qwen showed the highest validity rate (23/25). Its rate of items with three diagnostic distractors was also the highest, but remained low (4/25).

Eleven of Gemma’s 25 items failed the validity criterion, the highest failure rate among the three models, most often because no option was correct. Phi showed the lowest clarity rate (14/25) and accounted for all four non-viable outputs.

No inferential tests were conducted. These results are descriptive comparisons across five generations per model and task type.

Performance also varied markedly by task type (Table 3). The multi-step fraction condition showed the weakest descriptive performance across dimensions. Only 1 of 15 items had a correct key, 4 of 13 evaluable items met the intended demand, and none of the 13 evaluable items contained three diagnostic distractors. Rounding items, by contrast, all met the validity criterion (15/15), yet none contained three diagnostically valid distractors. This contrast illustrates that structural quality and instructional usefulness can diverge.

4.3 Failure Patterns and Readability

Table 4 summarizes the most frequent secondary failure patterns. No valid option and key–rationale inconsistency were the two most common failures. Model-specific clustering was also apparent: 11 of 14 key–rationale inconsistencies occurred in

Gemma, whereas all non-viable outputs, missing-rationale cases, and key–option mismatches occurred in Phi. These patterns suggest qualitatively different oversight demands rather than a simple difference in overall model quality.

Table 4: Recurring secondary failure patterns in generated items.

Failure type	<i>n</i>	Operational pattern
No valid option	16	No listed option was mathematically correct.
Key–rationale inconsistency	14	Stated reasoning did not produce the key; 11/14 were Gemma.
Non-viable output	4	Item could not be meaningfully evaluated; all were Phi.
Missing rationale	4	Required distractor rationale was absent; all were Phi.
Key–option mismatch	3	Designated key did not match a listed option; all were Phi.

Note. Failure flags are not mutually exclusive and may co-occur within an item.

As a descriptive readability check, Flesch–Kincaid estimates exceeded the nominal target grade for 39 of 75 items (52%). Median FK grade level was 1.5 for Qwen, 7.3 for Gemma, and 3.9 for Phi; Gemma also produced substantially longer items, with a median stem-and-option length of 43 words against 19 for Qwen. Because mathematical notation distorts readability formulas and stems were extracted automatically, FK was used descriptively rather than as a pass/fail criterion.

5 Discussion

This work contributes (1) a PD model positioning teachers as critical evaluators of AI-generated mathematics items, extending teacher-in-the-loop work (Reza et al., 2025) into teacher learning; (2) empirical evidence that answer-key accuracy alone is insufficient for judging item quality, with only 8 of 37 key-correct items containing three diagnostic distractors and only 7 of 75 items passing all six criteria; and (3) a measurement-grounded rubric linking AI fluency with assessment literacy. Multi-step fraction tasks were especially fragile, and model-specific failure profiles suggest different oversight demands.

All models were evaluated in out-of-the-box instruction-tuned configurations, providing a baseline for future adaptation. The readability findings identify one potential target for improvement: supervised fine-tuning, preference optimization, or reinforcement learning on grade-appropriate language could be investigated as ways to improve linguistic accessibility. Whether the observed failure patterns persist under alternative reasoning or adaptation settings also remains open.

These findings are descriptive and should be interpreted in light of the single-coder analysis, fixed prompt order, and absence of teacher implementation data. The study also covers only three models and one prompt realization per task type, and no formal coding-reliability estimate was obtained. Because task type was confounded with turn position within each session, task-level differences should not be interpreted as causal effects of task complexity.

Future work will test the productive-imperfection framework with in-service teachers following IRB approval.

References

- Lorin W. Anderson and David R. Krathwohl, editors. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.
- L. M. Desimone. 2009. [Improving impact studies of teachers' professional development: Toward better conceptualizations and measures](#). *Educational Researcher*, 38(3):181–199.
- S. E. Embretson and N. M. Kingston. 2018. [Automatic item generation: A more efficient process for developing mathematics achievement items?](#) *Journal of Educational Measurement*, 55(1):112–131.
- L. Favero, J. A. Pérez-Ortiz, T. Käser, and N. Oliver. 2025. [Leveraging small LLMs for argument mining in education: Argument component identification, classification, and assessment](#). *arXiv preprint arXiv:2502.14389*.
- Gemma Team. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- M. J. Gierl and T. M. Haladyna, editors. 2012. *Automatic Item Generation: Theory and Practice*. Routledge.
- C. S. Große and A. Renkl. 2007. [Finding and fixing errors in worked examples: Can this foster learning outcomes?](#) *Learning and Instruction*, 17(6):612–634.
- M. J. H. Masum, M. S. H. Majumder, M. M. Rahman, M. Z. Hasan, and F. Al Faisal. 2025. [Design and optimization of an offline AI model for enhancing student learning and academic assistance](#). In *2025 IEEE 7th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pages 1–6. IEEE.
- Microsoft. 2025. *Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras*. *arXiv preprint arXiv:2503.01743*.
- W. J. Popham. 2009. [Assessment literacy for teachers: Faddish or fundamental?](#) *Theory Into Practice*, 48(1):4–11.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Z. Reza, A. Mazur, M. T. Dugdale, and R. Ray-Chaudhuri. 2025. [Small models, big support: A local LLM framework for educator-centric content creation and assessment with RAG and CAG](#). *arXiv preprint arXiv:2506.05925*.
- F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, and S. Wang. 2025. [A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness](#). *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.

A Representative SLM-Generated Item

A.1 Generation Prompt

Prompt. You are writing a Grade 5 multiple-choice math question on adding two fractions with unlike denominators. Make it a multi-step word problem. Provide:

- the question
- four choices A–D (exactly one correct)
- the correct answer
- for each wrong choice, the student misconception it reflects

Use grade-appropriate numbers and wording.

A.2 Representative Phi-4-mini Output

Phi-4-mini, Run 1 — Multi-Step Fraction Addition

Question: Emily has $\frac{3}{4}$ of an apple pie slice left over from lunch yesterday; today she eats another third portion as dessert at dinner time. How much does this total fraction represent?

- A) 1
- B) $\frac{7}{12}$
- C) More than a whole (2 pieces)
- D) Cannot add them together without converting to improper fractions.

Correct Answer: B)

Misconceptions reflected by incorrect answers:

A) *Incorrectly summing numerators directly rather than finding common denominators first.*

B) Correct understanding of adding $\frac{3}{4} + \frac{1}{3} = (9+4)/12$ leading up to total as simplified fraction.

C) Miscalculating overestimation thinking that both fractions contribute beyond a singular whole without the proper addition process considered. Overestimating portions together leads to this misconception.

D) Assuming improper handling of mixed numerical types and not realizing direct methods can still apply using least common denominators for simple fractional sums.

Note. Model output is reproduced verbatim except for typographic formatting.

A.3 Researcher Evaluation

The designated answer is incorrect. The model's own procedure implies

$$\frac{3}{4} + \frac{1}{3} = \frac{9}{12} + \frac{4}{12} = \frac{13}{12},$$

rather than $\frac{7}{12}$. Thus, the response contains a direct inconsistency between the designated key and its own mathematical rationale. The exact value $\frac{13}{12}$ is not offered as an option. Option C correctly indicates that the sum exceeds one, but its parenthetical “2 pieces” is inaccurate; therefore, no listed option is wholly correct. The model nevertheless labels option C as a misconception. This example illustrates how a fluent SLM-generated item can contain multiple interacting defects requiring evaluation beyond answer-key accuracy alone.

B Rubric Grounding and Decision Rules

The six dimensions were designed to translate established principles from assessment literacy, automatic item generation, and cognitive-demand research into teacher-facing evaluation criteria.

D1 Key. Pass if the designated answer is mathematically correct.

D2 Validity. Pass if exactly one option is mathematically correct, the correct answer appears among the options, the key refers unambiguously to one listed option, options are distinct, and options contain response content only.

D3 Alignment. Pass if the item assesses the intended mathematical skill or concept.

D4 Demand. Pass if the reasoning required by the item matches the intended cognitive demand.

D5 Diagnosticity. Pass if all three designated distractors are reachable through identifiable student errors, match the stated misconception, and are supported by mathematically sound rationales.

D6 Clarity. Pass if the item is unambiguous, grade-appropriate, and sufficiently clear for the intended learner.