STEPwise: Diagnosing Students' Math Problem-Solving Skills via LLM-Generated Stepwise Sub-Questions

Yoonsu Kim[1], Hayeon Doh[2], Eunhye Kim[1], Jinho Son[3], and Juho Kim[1]

[1] School of Computing, KAIST, Daejeon, Republic of Korea
{yoonsu16,gracekim027,juhokim}@kaist.ac.kr
[2] Ewha Womans University, Seoul, Republic of Korea hydoh12@ewha.ac.kr
[3] AlgorithmLabs, Seoul, Republic of Korea sjhfam@algorithmlabs.co.kr

Abstract. Diagnosing students' mathematical problem-solving skills at scale is challenging due to the difficulty of eliciting fine-grained reasoning and constructing problem-specific rubrics. We present STEPwise, a teacher-in-the-loop system that supports the full workflow of process-level diagnosis, from diagnostic material generation to student assessment. STEPwise is designed to make students' reasoning visible at each problem-solving stage and support structured, scalable process-level assessment in mathematics classrooms. STEPwise is grounded in STEP (**ST**age-based **E**valuation of **P**roblem-solving skills), a diagnostic framework co-designed with six mathematics education experts that decomposes problem-solving into four stages and eight sub-skills. Given a math word problem, STEPwise automatically generates eight stage-specific sub-questions and corresponding three-tier rubrics using an LLM-based pipeline, which employs chain-of-thought problem solving, multi-verifier quality checking, and simulated student error generation. Teachers can review, edit, and provide feedback on all generated materials before deploying them to students. Once students submit responses, STEPwise scores them based on the rubrics and produces a diagnostic report summarizing each student's performance across reasoning stages.

Keywords: Mathematical problem-solving process, Process-level diagnosis, Large language models (LLMs), Diagnostic material generation

Demo Link: https://youtu.be/cyquqLOIEoo

System Presentation

Background and Motivation Diagnosing students' problem-solving processes in mathematics requires more than evaluating final answers—it demands close examination of how students reason at each stage. Yet in practice, teachers lack the tools to do this consistently at scale: without a shared diagnostic structure, stage-based assessments vary across problems, and constructing problem-specific rubrics for each stage is too costly to scale manually. STEPwise addresses both

challenges by automating the generation of diagnostic materials while keeping teachers in the loop for review and refinement.

The STEP Framework STEPwise is grounded in STEP (STage-based Evaluation of Problem-solving skills), a diagnostic framework co-designed with six mathematics education experts through iterative interviews. STEP decomposes mathematical problem-solving into four stages—Problem Comprehension, Information Structuring, Mathematical Expression, and Mathematical Computation—each articulated with two diagnostically distinct sub-skills, yielding eight sub-skills in total. Each sub-skill is mapped to primary error types drawn from Newman's error analysis (e.g., Reading Error, Comprehension Error, Transformation Error) [1], providing a theoretically grounded structure for both question design and rubric construction.

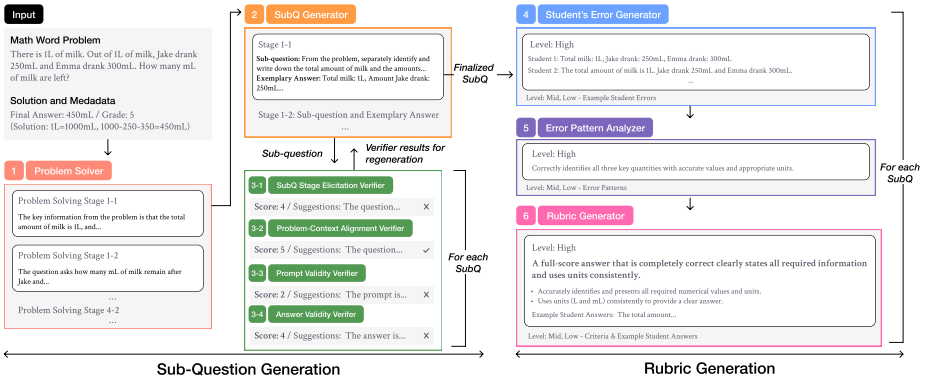

Fig. 1. The LLM-based Pipeline for STEPwise.

LLM-based Pipeline for Diagnostic Material Generation Given a math word problem, STEPwise generates diagnostic materials through a multi-stage LLM pipeline consisting of two main components.

Sub-Question Generation. The pipeline begins with a Problem Solver (Figure 1-1) that reasons through the problem step by step via chain-of-thought prompting [2] guided by the STEP framework, and classifies the problem into a mathematical content domain (e.g., Numbers and Operations, Geometry and Measurement) to apply domain- and grade-specific curriculum constraints. A Sub-Question Generator (Figure 1-2) then transforms the identified reasoning steps into eight stage-specific sub-questions, one per STEP sub-skill, each targeting a single reasoning step. To improve educational soundness, each sub-question undergoes a **generate–verify–regenerate cycle** assessed by a multi-verifier module [3] with four verifiers: *stage elicitation* (Figure 1-3-1), *context alignment* (3-2), *prompt validity* (3-3), and *answer validity* (3-4). Each

verifier scores on a 0–5 scale, and sub-questions scoring below the threshold are regenerated with verifier feedback for up to three iterations.

Rubric Generation. For each sub-question, a Student Error Generator (Figure 1-4) simulates nine student responses across three performance levels, drawing on Newman's error types [1] and grade-level achievement standards to ensure realistic, level-appropriate responses. An Error Pattern Analyzer (Figure 1-5) then extracts distinct, non-overlapping error patterns and establishes clear boundaries across levels. Finally, a Rubric Generator (Figure 1-6) constructs a three-tier rubric—High (correct), Mid (partial understanding), and Low (fundamental error)—with level-specific criteria and representative examples grounded in the simulated responses.

System Interface and Teacher Workflow STEPwise supports the full diagnostic workflow through four phases: (1) Material Generation: Teachers input a math word problem; the system generates an 8-step solution process, subquestions, and rubrics sequentially. (2) Teacher Review and Editing: At each stage, teachers can directly edit outputs, provide natural language feedback to trigger targeted regeneration before approving and proceeding to the next stage. This teacher-in-the-loop design ensures curriculum alignment and pedagogical appropriateness without requiring full manual authoring. (3) Student Deployment: Finalized materials are exported as a PDF worksheet for classroom distribution. (4) Diagnosis and Reporting: After students submit responses, STEPwise scores each response against the rubrics using an LLM and generates a diagnostic report summarizing performance by problem and by problem-solving stage, allowing teachers to pinpoint where students' reasoning breaks down.

Unlike existing LLM-based approaches that focus on standalone problem generation or final-answer scoring, STEPwise covers the *full pipeline of process-level diagnosis*, integrating question generation, rubric construction, automated scoring, and stage-level reporting into a single coherent system. The *teacher-in-the-loop design* balances automation with pedagogical oversight, enabling teachers to refine materials without full manual authoring. Finally, STEPwise is *problem-agnostic and curriculum-adaptable*, instantiating STEP for any math word problem across different national curriculum standards.

References

1. Newman, M.A.: An analysis of sixth-grade pupil's error on written mathematical tasks. Victorian Institute for Educational Research Bulletin **39**, 31–43 (1977)
2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
3. Worden, E., Vanacore, K., Haim, A., Heffernan, N.: Scaling effective ai-generated explanations for middle school mathematics in online learning platforms. In: Proceedings of the Twelfth ACM Conference on Learning @ Scale. p. 40–49. L@S '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3698205.3729546