

MALRULELIB: Large-Scale Executable Misconception Reasoning with Step Traces for Modeling Student Thinking in Mathematics

Xinghe Chen*
Rice University
xc42@rice.edu

Naiming Liu*
Rice University
nl35@rice.edu

Shashank Sonkar
University of Central Florida
shashank.sonkar@ucf.edu

Abstract

Student mistakes in mathematics are often systematic: a learner applies a coherent but wrong procedure and repeats it across contexts. We introduce MALRULELIB, a learning-science-grounded framework that translates documented misconceptions into executable procedures, drawing on 67 learning-science and mathematics education sources, and generates step-by-step traces of malrule-consistent student work. We formalize a core student-modeling problem as **Malrule Reasoning Accuracy (MRA)**: infer a misconception from one worked mistake and predict the student’s next answer under cross-template rephrasing. Across nine language models (4B–120B), accuracy drops from 66% on direct problem solving to 40% on cross-template misconception prediction. MALRULELIB encodes 101 malrules over 498 parameterized problem templates and produces paired dual-path traces for both correct reasoning and malrule-consistent student reasoning. Because malrules are executable and templates are parameterizable, MALRULELIB can generate over one million instances, enabling scalable supervision and controlled evaluation. Using MALRULELIB, we observe cross-template degradations of 10–21%, while providing student step traces improves prediction by 3–15%. We release MALRULELIB as infrastructure for educational AI that models student procedures across contexts, enabling diagnosis and feedback that targets the underlying misconception. MALRULELIB code is available [here](#).

1 Introduction

A student who computes $\frac{1}{2} + \frac{1}{3} = \frac{2}{5}$ is not guessing. They are applying a coherent but flawed procedure. Add the numerators, add the denominators. Learning scientists have shown that many mathematical errors arise from such systematic procedures.

*Equal contribution.

Model	CRA	MRA	Forward MRA
Llama-3.3-70B	70.4%	34.6%	39.6%
Qwen3-80B-Think	70.1%	56.4%	53.9%
gpt-oss-120b	65.0%	56.9%	48.8%
<i>Average</i>	68.5%	49.3%	47.5%

Table 1: Three evaluation settings for the educational Turing Test. **CRA**: solve problems correctly. **MRA**: infer misconceptions from examples and predict student answers. **Forward MRA**: receive explicit misconception descriptions and predict answers.

They are often called *malrules*, misconceptions, or procedural bugs, and they are stable, diagnosable, and instructionally meaningful (Brown and Burton, 1978a; Siegler et al., 2012). For a tutor, the key step is not verifying correctness. It is inferring *which* malrule a student is using and predicting how it will reappear on the next problem, so feedback targets the underlying reasoning.

Can modern AI do this? Sonkar et al. (2025) propose a direct “Educational Turing Test” for student modeling. Given evidence of a student’s misconception, can a model predict the specific errors that student will make on new problems? We operationalize this test with **Malrule Reasoning Accuracy (MRA)**. In MRA, a model is shown a student’s incorrect solution to one problem. The model must infer the underlying malrule from that example and then predict the student’s answer on a new problem.

Critically, the new problem often changes surface form. Table 2 illustrates the challenge. The model sees a student evaluate $\sqrt{x^2 + 25}$ at $x=8$ and answer 13, which implies the malrule $\sqrt{a^2 + b^2} = a + b$. The model must then apply the same malrule to a different template, such as a distance word problem that implicitly requires $\sqrt{8^2 + 3^2}$. This cross-template generalization is what tutoring demands. Students rarely repeat the same question, but they do repeat the same miscon-

Forward MRA	MRA (Cross-Template)
<p><i>System:</i> You are simulating a student who has a specific mathematical misconception. Apply the described misconception consistently to solve the problem.</p> <p><i>User:</i> A student has the following misconception: Students distribute square root over addition: $\sqrt{a^2 + b^2} = a + b$ Apply this misconception to solve: Evaluate $f(x) = \sqrt{x^2 + 4}$ when $x = 3$. What is $f(3)$?</p>	<p><i>System:</i> You are an expert in identifying and understanding student mathematical misconceptions. Given an example of a student’s incorrect answer, identify the systematic error and apply it to predict answers for new problems.</p> <p><i>User:</i> A student solved this problem incorrectly: Problem: Evaluate $f(x) = \sqrt{x^2 + 25}$ when $x = 8$. Student’s Answer: 13 Now predict what this same student would answer for: You walk 8 blocks east and 3 blocks north. What is the straight-line distance from your starting point?</p>
<p><i>Expected:</i> 5 (correct: $\sqrt{13} \approx 3.61$)</p>	<p><i>Expected:</i> 11 (correct: $\sqrt{73} \approx 8.54$)</p>

Table 2: Prompts for Forward MRA and MRA tasks. **Forward MRA** provides an explicit misconception description; the model must translate it into procedural errors. **MRA** provides only a worked example; the model must infer the misconception pattern and generalize it to a new problem format (here, from algebraic to word problem). Both prompts target the same underlying malrule: distributing square roots over addition.

ception across contexts.

Our results expose a sharp gap between doing mathematics and modeling student thinking. On three representative large models (70B–120B) shown in Table 1, models achieve 68.5% accuracy on direct problem solving (**CRA**: Correct Reasoning Accuracy), but only 49.3% on **cross-template** misconception prediction from an example (**MRA**). **Mathematical reasoning ability does not transfer to student modeling**. We also evaluate **Forward MRA**, where the model is given an explicit natural-language description of the misconception and must apply it, and find performance remains limited (47.5% on the same models). Table 5 reports the full benchmark across all nine models and experimental settings.

Misconceptions are well studied, but the field lacks infrastructure that treats them as computational objects. Most resources describe misconceptions, but they do not operationalize them as procedures that can be executed across many templates with malrule-consistent intermediate steps (Lucy et al., 2024). This makes it difficult to generate training data, run controlled evaluations, or measure cross-template student modeling at scale.

We address this gap with MALRULELIB, a learning-science-grounded framework that encodes misconceptions as executable procedures and pairs them with diverse problem templates. For each instantiated problem, MALRULELIB generates dual-path solution traces: a fully correct solution and a malrule-consistent student solution, both with step-by-step work. Because malrules are executable and templates are parameterizable, MALRULELIB can generate large-scale data with ground-truth malrule identity and trace-level supervision. Table 3 illus-

trates the structure. Each malrule is shown on two different templates, often shifting from a symbolic expression to a word problem. The student’s work remains systematically consistent with the same underlying procedure, even as surface features change. This is the core challenge for personalized learning and for our benchmark: models must infer the malrule from evidence and predict misconception-consistent work under cross-template shifts, not merely reproduce a template-specific error pattern.

We make three contributions:

1. **A learning-science-grounded misconception library as executable procedures.** We translate 101 documented malrules into computational objects: each misconception is implemented as an executable procedure and paired with 498 diverse problem templates spanning 22 mathematical categories. Each malrule is grounded in the learning-science literature, drawing from 67 papers (Appendix Tables 8–11). This grounding ensures the benchmark targets real, instructionally meaningful error patterns and supports downstream tutoring actions beyond prediction.
2. **Dual-path solution traces at scale.** For every template instance, we generate aligned step-by-step work for *both* correct reasoning and malrule-consistent student reasoning. Because templates are parameterizable and malrules are executable, MALRULELIB can generate millions of such paired instances with dual traces, providing trace-level supervision for training and controlled evidence for evaluation. In our experiments, supplying student steps yields substantial improvements in mis-

Category	Malrule	Problem	Student's Work
Radicals	$\sqrt{a^2 + b^2} = a + b$	T1: Evaluate the function $f(x) = \sqrt{x^2 + 25}$ when $x = 8$. What is $f(8)$? T2: You walk 8 blocks east and 3 blocks north. What is the straight-line distance from your starting point?	$\sqrt{x^2} + \sqrt{25} = x + 5; 8 + 5 = \mathbf{13}$ $\sqrt{8^2} + \sqrt{3^2} = 8 + 3 = \mathbf{11}$
Order of Operations	Addition before subtraction	T1: Evaluate: $29 - 28 + 12$ T2: Starting at 45°F , the temperature decreases by 5°F , then increases by 3°F . What's the result?	$28 + 12 = 40; 29 - 40 = -\mathbf{11}$ $5 + 3 = 8; 45 - 8 = \mathbf{37^\circ\text{F}}$
Functions	$f(a + b) = f(a) + f(b)$	T1: Given $f(x) = x^3$, evaluate $f(11 + 10)$ T2: Given $f(x) = x + 3 $, evaluate $f(8 + 4)$	$f(11) = 1331, f(10) = 1000; 1331 + 1000 = \mathbf{2331}$ $f(8) = 11, f(4) = 7; 11 + 7 = \mathbf{18}$
Division	Larger \div smaller always	T1: 4 cookies are shared equally among 6 children. How much does each child get? T2: 4 meters of ribbon is divided into 5 equal strips. How long is each strip in meters?	$4 \div 6 \rightarrow 6 \div 4 = \mathbf{1.5}$ $4 \div 5 \rightarrow 5 \div 4 = \mathbf{1.25}$
Subtraction	Borrow without decrementing	T1: Calculate: $408 - 384$ T2: A store had 561 items in stock. After selling 526 items, how many remain?	$8 - 4 = 4, 10 - 8 = 2, 4 - 3 = 1 \rightarrow \mathbf{124}$ $11 - 6 = 5, 6 - 2 = 4, 5 - 5 = 0 \rightarrow \mathbf{45}$

Table 3: Examples of malrules from MALRULELIB, showing two templates per misconception. Each malrule produces systematic errors across different problem formats: algebraic expressions and word problems. The template diversity illustrates the cross-template generalization challenge: models must recognize the same underlying misconception despite surface-level differences. See Appendix Table 7 for additional examples.

conception prediction, validating the value of step-level data for student modeling.

- The first large-scale benchmark for cross-template misconception prediction.** We evaluate 9 language models (4B–120B) on **Malrule Reasoning Accuracy** under controlled conditions that separate same-template from cross-template generalization and compare answer-only versus with-steps evidence. Across models, cross-template performance drops by 10–21 points, and step evidence produces consistent gains of 3–15 points, exposing a persistent gap between problem solving and modeling misconception-driven student behavior.

2 Related Work

If you can both listen to children and accept their answers not as things to just be judged right or wrong but as pieces of information which may reveal what the child is thinking you will have taken a giant step toward becoming a master teacher rather than merely a disseminator of information.

— Easley and Zwoyer (1975)

This “teaching by listening” view motivates a diagnostic stance toward student errors. The goal is not only to mark an answer right or wrong, but to infer the underlying procedure that produced it. Brown and Burton (1978a) operationalized this idea in the BUGGY line of work. They modeled systematic errors as small, structured edits to a correct procedure, and used these diagnostic models to explain errors, predict future mistakes, and even generate diagnostic tests. MALRULELIB follows this tradition and extends it to a broader range of mathematical domains and to modern evaluation of language models.

2.1 The BUGGY Tradition: Procedural Misconceptions in Learning Science

The study of mathematical misconceptions has deep roots in cognitive science, but BUGGY marked an important shift in emphasis. Rather than treating wrong answers as noise, Brown and Burton (1978a) argued that many errors are coherent, rule-governed procedures. In their framing, a child writing $\frac{1}{2} + \frac{1}{3} = \frac{2}{5}$ is not confused about addition. The child is overgeneralizing whole-number procedures to fractions. This procedural view transformed how educators and tutoring systems reason about student knowledge. The target of instruction becomes the underlying rule, not the surface error.

Subsequent decades uncovered misconceptions across mathematical domains. Behr et al. (1984) and Ni and Zhou (2005c) documented “whole number bias” in fraction arithmetic, where students treat $\frac{a}{b}$ as two independent numbers rather than a single quantity. Resnick et al. (1989b) identified systematic errors in decimal comparison, such as believing $0.29 > 0.3$ because 29 is greater than 3. Matz (1980) catalogued algebraic misconceptions, and Vlassis (2004a) traced difficulties with signed numbers to overgeneralized subtraction rules.

A key finding across this literature is that misconceptions are remarkably stable. Siegler et al. (2012) show that fraction magnitude understanding in middle school predicts mathematical achievement years later. Once formed, faulty mental models resist correction. This stability makes misconception prediction both tractable and educationally valuable. MALRULELIB operationalizes this literature as executable procedures. It encodes malrules along with prevalence information, root-cause hypotheses, and remediation strategies, then uses them to generate structured student work.

2.2 Misconception Datasets and Benchmarks

Prior work has developed resources for studying student errors, but existing datasets share common limitations. ASSISTments (Heffernan and Heffernan, 2014; Worden et al., 2026) logs student interactions at scale and final answers, but they do not capture actual student step-by-step solutions. More recent work has begun addressing the rationale gap. Sonkar et al. (2024a) introduce MALALGOQA, a dataset of roughly 807 mathematics comprehension questions, each annotated with misconceptions. They find that LLMs exhibit substantial drops when identifying misconceptions compared to correct-answer rationales, foreshadowing our CRA–MRA gap. However, MalAlgoQA provides static rationales curated from existing items rather than executable misconception procedures grounded in learning-science research. It also evaluates questions largely in isolation, without controlled cross-template generalization tests of the same misconception across surface forms.

2.3 AI for Education and Student Modeling

Student modeling is central to intelligent tutoring systems, such as cognitive tutors (Anderson et al., 1995; Liu et al., 2023), LLM based tutors (Sonkar et al., 2023, 2024b) and knowledge tracing (Corbett and Anderson, 1994; Sonkar et al., 2020; Sonkar

and Baraniuk, 2023). These approaches work well in narrow domains but require substantial hand engineering. Sonkar et al. (2025) argue that educational AI should be evaluated by whether it can predict student behavior, a Turing-like test for personalized education. Our work provides the first large-scale benchmark for this capability. In doing so, we connect modern LLM evaluation to the earlier diagnostic modeling tradition exemplified by BUGGY.

3 The MALRULELIB Framework

Building AI that understands student thinking requires data that captures how students actually reason when making mistakes. We need not just wrong answers, but the cognitive processes behind them. MALRULELIB is a Python framework designed to generate such data at scale, grounded in decades of learning science research (see Table 8).

3.1 Design Principles

Three principles guided our framework design.

First, **learning science grounding**. Every malrule in MALRULELIB traces to documented research on student misconceptions. We don’t invent plausible errors; we encode errors that real students consistently make, complete with prevalence data and cognitive explanations. This grounds the framework in empirical findings rather than speculation about what might confuse students. A full mapping of malrules to their academic sources appears in Appendix Tables 8–11.

Second, **cognitively faithful student solutions**. A critical design decision: for every problem, we generate step-by-step solutions showing *both* correct and incorrect reasoning paths. The malrule path captures how students actually think when applying a misconception—not just wrong answers, but the cognitive process behind them. This dual-path generation enables training LLMs to understand and predict student reasoning, not merely evaluate correctness.

Third, **template diversity**. Each malrule is instantiated through multiple templates: basic formulations, structural variants, real-world contexts, and word problems. The 498 templates (4.9 per malrule on average) enable rigorous cross-template generalization testing—assessing whether models truly understand misconceptions or merely pattern-match surface features.

Category	M	T	Description
NUMBER & OPERATIONS (54 malrules, 277 templates)			
Whole Number Ops	16	97	Place value, regrouping
Fractions & Ratios	13	63	Part-whole, proportions
Decimals & Percents	16	83	Notation, conversions
Signed Numbers	9	34	Integers, absolute value
ALGEBRA (37 malrules, 168 templates)			
Exponents & Radicals	12	72	Laws of exponents, roots
Expressions & Equations	21	84	Simplifying, PEMDAS
Functions	4	12	Notation, input-output
GEOMETRY & MEASUREMENT (8 malrules, 20 templates)			
Geometry	6	14	Area, perimeter, volume
Coordinate Geometry	2	6	Ordered pairs, graphing
DATA & MODELING (4 malrules, 33 templates)			
Data & Word Problems	4	33	Statistics, translation

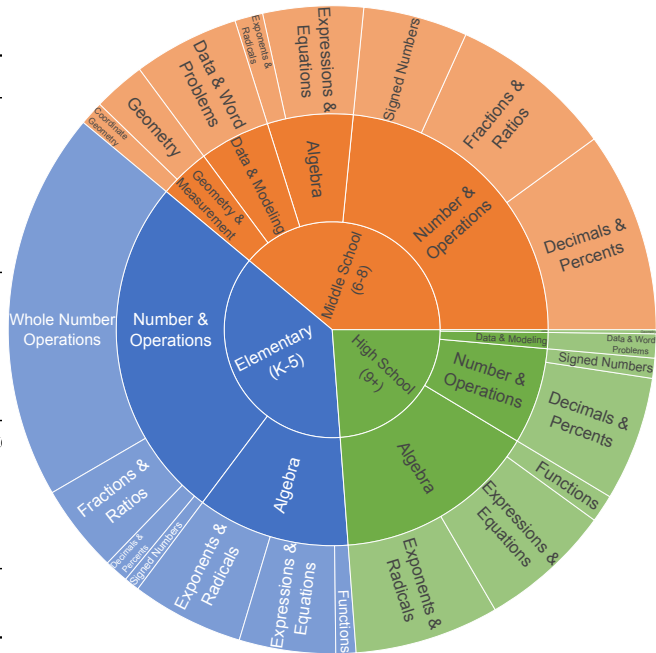


Figure 1: MalruleLib framework overview. **Left:** Classification by NCTM Content Strand (M=Malrules, T=Templates). **Right:** Distribution by developmental stage showing coverage across grade levels.

Malrule Architecture MALRULELIB is open-source and designed for extensibility (code is available at <https://github.com/sonkar-lab/malrulelib>). Each malrule is a self-contained module with four components:

```
[category]/[malrule_name]/
|--problem_generator.py # Templates
|--correct_algorithm.py # Correct steps
|--malrule_algorithm.py # Malrule steps
|--test_malrule.py # Unit tests
```

The **problem generator** defines multiple templates—parameterized problem structures that produce diverse instances. The **correct algorithm** and **malrule algorithm** implement the mathematically sound and misconception-based procedures respectively, each generating step-by-step reasoning. This modular design makes it straightforward to add new malrules: implement the four components, and the framework handles problem generation, validation, and integration.

3.2 Coverage Statistics

MALRULELIB currently encodes **101 malrules** spanning common student errors from elementary through early high school mathematics, organized into **22 mathematical categories** aligned with 10 NCTM content strands (Figure 1). These malrules are instantiated through **498 templates** (4.9 per malrule on average), enabling generation of thousands of unique problem instances with **dual-path**

step generation for every problem. Categories range from elementary arithmetic (whole number operations, basic fractions) through middle school topics (exponents, linear equations) to early algebra (factoring, functions). Table 6 shows the distribution across NCTM strands.

Deep analysis of these 498 templates reveals intentional pedagogical design along two dimensions grounded in learning science.

Context domain diversity. Templates span 10 context domains: 63.3% use abstract mathematical notation while 36.7% embed problems in real-world contexts such as measurement (11.0%), money (9.2%), time/distance (3.8%), science (3.0%), sports (3.0%), and food (3.0%). This design reflects research on *transfer of learning*: students often fail to apply knowledge across contexts, exhibiting “inert knowledge” that activates only in familiar settings (Bransford et al., 1999). A student may add fractions correctly in abstract form but apply the malrule when the same problem appears as pizza slices. By varying context while holding the misconception constant, we assess whether models exhibit similar context-dependent failures.

Scaffolded complexity. Templates follow a pedagogical progression: *basic* (18.5%), *variant* (50.8%), *context* (6.2%), and *word problem* (24.5%). This scaffolding enables assessment across difficulty gradients. Word problems par-

ticularly test whether misconceptions persist under additional cognitive load from reading comprehension. Additional template coverage details appear in the Appendix D.

3.3 A Misconception Generator at Million-Instance Scale

MALRULELIB is not just a collection of misconception labels. It is a *generator* that can produce over one million distinct problem instances with different parameters and surface forms. The core idea is simple: if misconceptions are stable procedures, they should be encoded as *executable programs*. Once a malrule is executable, we can systematically generate both (i) the correct solution trace and (ii) the misconception-consistent student trace, at scale and under tight control.

Templates are parameterized by grade-banded value ranges and difficulty presets, enabling both systematic coverage and pedagogically meaningful distributions. We estimate generation capacity by counting valid parameter assignments and template variants that satisfy the constraints for a chosen grade and difficulty setting yielding over one million distinct instances. This scale and mechanism makes it feasible to fine-tune or instruction-tune models directly on executed misconception traces. Additionally, each template exposes a small set of parameters together with malrule-specific constraints that guarantee the misconception is actually triggered. For borrowing-related subtraction, for example, MALRULELIB constructs operands digit by digit and enforces the inequalities that force borrowing at the intended place value.

4 Benchmark Design

Using MALRULELIB, we construct a benchmark that operationalizes the Turing Test for educational AI: can models predict student errors well enough to demonstrate understanding of student reasoning?

4.1 Task Definitions

We introduce notation to formally define our evaluation tasks. Let $m \in \mathcal{M}$ denote a malrule from set of 101 malrules. Each malrule has an associated set of templates $\mathcal{T}_m = \{t_1, t_2, \dots\}$. An instance $i \sim t$ is a concrete problem sampled from template t by instantiating its parameters with specific values. For each instance i , we denote $a_c(i)$ and $a_m(i)$ as the correct and malrule answers respectively, and $S_m(i)$ as the step-by-step malrule reasoning.

Malrule Reasoning Task. Given a source instance i_s with its malrule answer $a_m(i_s)$, and a target problem i_t , the model must predict $a_m(i_t)$, the answer this student would give if applying the same malrule. We vary two factors: *template condition* and *prompt condition*. For template condition, *same-template* samples source and target from the same template ($i_s, i_t \sim t$), while *cross-template* samples from different templates ($i_s \sim t_1, i_t \sim t_2$ where $t_1 \neq t_2$). For prompt condition, *answer-only* provides the source problem and malrule answer, while *with-steps* additionally provides the malrule reasoning steps $S_m(i_s)$. Table 4 summarizes the four experimental conditions.

Condition	Predict
<i>Same-template: $i_s, i_t \sim t$</i>	
Answer-only: $(i_s, a_m(i_s), i_t)$	$a_m(i_t)$
With-steps: $(i_s, a_m(i_s), S_m(i_s), i_t)$	$a_m(i_t)$
<i>Cross-template: $i_s \sim t_1, i_t \sim t_2$</i>	
Answer-only: $(i_s, a_m(i_s), i_t)$	$a_m(i_t)$
With-steps: $(i_s, a_m(i_s), S_m(i_s), i_t)$	$a_m(i_t)$

Table 4: MRA experimental conditions.

Forward MRA Task. We also evaluate whether models can apply a misconception given its description. Let $D(m)$ denote a natural language description of malrule m . Given $D(m)$ and a problem i , the model must predict $a_m(i)$. Table 2 illustrates both MRA and Forward MRA prompt formats.

Correct Reasoning Accuracy (CRA). To contextualize the difficulty of MRA and Forward MRA, we also evaluate whether models can solve problems correctly. Given a problem i , the model must predict $a_c(i)$, the correct answer.

4.2 Sampled Dataset Statistics

We sample ~ 10 instances per template from all 101 malrules, yielding 4,991 problem instances across 498 (malrule, template) groups. For *same-template* pairs, we sample (i_s, i_t) from the same group, selecting up to 10 pairs per group. For *cross-template* pairs, we sample $i_s \sim t_1$ and $i_t \sim t_2$ where $t_1 \neq t_2$, selecting 100 pairs per malrule (77 malrules have $|\mathcal{T}_m| \geq 2$). This yields 12,706 pairs: 5,006 same-template and 7,700 cross-template.

Inference calls. For MRA, each of the 12,706 pairs is evaluated under two prompt conditions (answer-only and with-steps), yielding ~ 25 K calls per model. For CRA, each of the 4,991 problem

Model	CRA	FMRA		MRA (no steps)				MRA (w/ steps)			
		Acc	Δ	Same	Δ	Cross	Δ	Same	Δ	Cross	Δ
gpt-oss-20b	65.2	46.5	-18.7	72.7	+7.5	53.6	-11.6	78.5	+13.2	57.7	-7.6
Qwen3-4B	67.3	17.4	-49.9	61.1	-6.2	39.1	-28.2	55.7	-11.6	34.6	-32.7
Phi-4	64.9	37.4	-27.5	50.0	-14.9	36.7	-28.2	64.9	-0.0	47.2	-17.7
Phi-4-mini	55.2	9.3	-45.9	26.7	-28.5	18.2	-37.0	42.5	-12.7	26.3	-28.9
Llama-3.1-8B	63.8	6.7	-57.1	25.3	-38.5	17.7	-46.1	43.6	-20.1	29.4	-34.4
<i>Small avg</i>	63.3	23.5	-39.8	47.2	-16.1	33.1	-30.2	57.1	-6.2	39.0	-24.3
gpt-oss-120b	65.0	48.8	-16.1	77.1	+12.1	56.9	-8.1	81.5	+16.5	60.9	-4.1
Qwen3-80B-Think	70.1	53.9	-16.2	74.2	+4.1	56.4	-13.7	77.2	+7.1	59.8	-10.3
Qwen3-80B-Inst	69.8	30.6	-39.2	69.9	+0.1	51.3	-18.5	73.2	+3.4	54.4	-15.4
Llama-3.3-70B	70.4	39.6	-30.7	47.7	-22.7	34.6	-35.8	64.4	-5.9	48.5	-21.9
<i>Large avg</i>	68.8	43.2	-25.6	67.2	-1.6	49.8	-19.0	74.1	+5.3	55.9	-12.9
Overall	65.7	32.3	-33.5	56.1	-9.7	40.5	-25.3	64.6	-1.1	46.5	-19.2

Table 5: Performance across experimental conditions. CRA = problem solving. FMRA = applying described misconception. MRA = predicting student answer from example. Same/Cross = template generalization. Δ = gap from CRA (negative = below CRA). Models grouped by size, sorted by Cross MRA. Across models, CRA exceeds MRA by large margins, and the gap widens under cross-template evaluation. Providing step-by-step reasoning traces typically improves MRA.

instances is solved once, yielding ~ 5 K calls. For Forward MRA, each instance is evaluated with its malrule description, yielding another ~ 5 K calls. In total, each model requires ~ 35 K inference calls. With 9 models evaluated, the benchmark comprises ~ 320 K total calls.

4.3 Models and Evaluation

We evaluate nine language models spanning two size categories. Large models (70–120B parameters) are gpt-oss-120b (Agarwal et al., 2025), Qwen3-80B-Think (Yang et al., 2025), Qwen3-80B-Instruct, and Llama-3.3-70B (Touvron et al., 2023). Small models (4–20B parameters) are gpt-oss-20b, Qwen3-4B, Phi-4 (Abdin et al., 2024), Phi-4-mini, and Llama-3.1-8B. Following model card recommendations, we use the following sampling parameters: for Qwen3 models in thinking mode, temperature 0.6, top- p 0.95, and top- k 20; for Qwen3 in non-thinking mode, temperature 0.7 and top- p 0.8; and for gpt-oss models, temperature 1.0 and top- p 1.0. Our primary metric is **Malrule Reasoning Accuracy (MRA)** to measure performance of malrule reasoning task: whether the model predicts the specific wrong answer produced by the malrule. We use normalized matching for algebraic expressions and numerical matching with tolerance for decimal answers.

5 Results and Discussion

Table 5 summarizes performance on three tasks: **CRA** (solve correctly), **MRA** (infer a student’s malrule from one example and predict the next answer), and **Forward MRA (FMRA)** (apply a described misconception). We additionally separate **Same** versus **Cross** template evaluation to measure generalization, and we compare **answer-only** versus **with-steps** evidence to quantify the value of reasoning traces. Figure 2 and Table 6 break results down by mathematical domain.

5.1 Capability gaps in student modeling

MRA remains below CRA. Models solve problems far better than they predict misconception-driven answers. Overall CRA is 65.7%, while cross-template MRA is only 40.5% (answer-only) and 46.5% (with steps), leaving gaps of 25.3 and 19.2 points (Table 5). This gap captures a core barrier for tutoring: competence at producing correct mathematics does not imply competence at modeling how a student will systematically be wrong.

A key reason is that MRA is a different computation than CRA. CRA is a single forward pass toward correctness, whereas MRA requires both inferring a latent procedure (the student’s malrule) from limited evidence and then executing that same flawed procedure on a new instance. This counterfactual execution conflicts with strong training priors for correct, helpful answers and is consistent

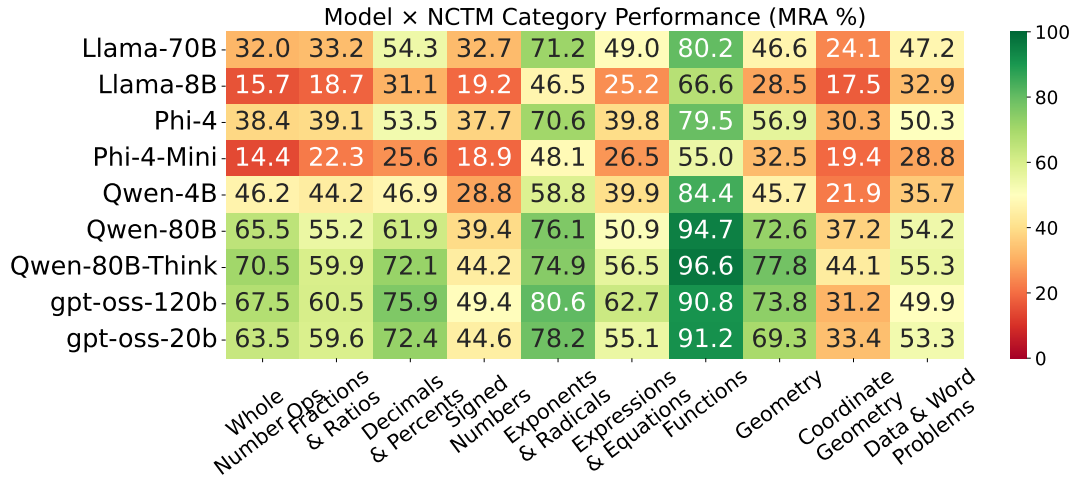


Figure 2: Performance by mathematical category. Functions is easiest (82%), Coordinate Geometry hardest (29%).

Category	Strand	MRA
Whole Number Ops	Number & Operations	46.0
Fractions & Ratios	Number & Operations	43.6
Decimals & Percents	Number & Operations	54.9
Signed Numbers	Number & Operations	35.0
Exponents & Radicals	Algebra	67.2
Expressions & Equations	Algebra	45.1
Functions	Algebra	82.1
Geometry	Geometry & Measurement	56.0
Coordinate Geometry	Geometry & Measurement	28.8
Data & Word Problems	Data & Modeling	45.3

Table 6: Average MRA by NCTM category across all models.

with a supervision imbalance: models see extensive curated data for correct reasoning during pretraining and instruction tuning, but far less curated data where an incorrect procedure is executed consistently and treated as the intended output.

Finally, note that same-template MRA can exceed CRA for some models, indicating that reproducing a local error pattern within a template family can be easier than solving the underlying mathematics. This is another reason cross-template evaluation is essential.

FMRA remains below CRA. Forward MRA is substantially below CRA: overall FMRA is 32.3% versus CRA at 65.7% (Table 5). In FMRA, the misconception is explicitly stated, so the failure is not lack of access to the rule. Rather, it reflects difficulty converting a natural-language description into a faithful, repeatable algorithmic transformation. Forward MRA requires translating a high-level misconception statement into a concrete sequence of algebraic operations. Phrases such as “distribute,” “cancel,” or “add inside” require the model to decide exactly where and how the flawed operation applies. In addition, instruction tuning encourages models to correct misconceptions rather than simu-

late them, creating a tension between being correct and behaving like a mistaken student.

Examples outperform descriptions. Despite giving the misconception explicitly, FMRA is generally worse than example-based MRA: overall FMRA is 32.3%, while cross-template MRA reaches 40.5 to 46.5% depending on whether steps are provided (Table 5). This pattern suggests that examples provide operational semantics that align with LLM strengths. A single worked mistake implicitly defines the mapping from problem structure to erroneous output, enabling in-context induction of the malrule. This suggests that executed examples with solution traces may be a more effective supervision format than descriptions alone for improving misconception application.

5.2 Value of step evidence and generalization

Reasoning traces help most models. MALRULELIB provides full step-by-step solution traces for both correct reasoning and malrule-based student reasoning, and these traces translate into measurable gains in student modeling. On average, cross-template MRA rises from 40.5% to 46.5% when steps are included, and many individual mod-

els see improvements (Table 5). This result validates our design decision to generate dual-path reasoning traces. It also points to a concrete training direction: supervised fine-tuning on malrule-consistent step traces can teach models to execute them faithfully under cross-template shifts.

Cross-template generalization is the bottleneck.

Another core contribution of MALRULELIB is that each malrule is instantiated across many diverse templates, enabling a direct test of whether models represent misconceptions as abstract procedures rather than template cues. Under this cross-template setting, performance drops sharply. Overall, same-template MRA (answer-only) is 56.1% while cross-template is 40.5%. With steps, the same-template score is 64.6% while cross-template is 46.5% (Table 5). Every model degrades when the surface form changes, indicating that many same-template successes can be achieved by template-level pattern matching rather than an abstract representation of the malrule. For example, cross-template items include shifts where the misconception must be applied inside a different wrapper, such as moving from a direct radical simplification to a distance word problem that implicitly requires $\sqrt{a^2 + b^2}$. These are precisely the settings where a tutor must recognize and anticipate the same flawed procedure across contexts.

Large domain spread. Performance varies sharply by mathematical category. Functions is easiest (82.1%) while Coordinate Geometry is hardest (28.8%), with Signed Numbers also low (35.0%), yielding a 53-point spread (Table 6). Figure 2 shows this pattern is consistent across model families. The spread in Table 6 suggests that student modeling is not monolithic and should be validated by domain before deployment. In practice, tutoring systems may require fine-tuning for certain categories such as geometry and signed arithmetic.

6 Conclusion

We introduced MALRULELIB, a learning-science-grounded framework that encodes 101 documented mathematical malrules over 498 problem templates and generates dual-path step-by-step traces for both correct reasoning and malrule-consistent student reasoning. Using MALRULELIB, we built the first large-scale benchmark for misconception-based student modeling with controlled same-template versus cross-template evaluation. Across nine

language models (4B–120B), we find a persistent gap between problem solving and predicting misconception-driven answers, with substantial cross-template degradations. Providing student work yields consistent gains, motivating evaluation protocols that test cross-context generalization and motivating educational systems to capture intermediate reasoning, not only final answers. Because malrules are executable and templates are parameterizable, MALRULELIB can generate large-scale training data with malrule-consistent intermediate steps for fine-tuning and instruction tuning. We release MALRULELIB as infrastructure for student modeling, a core capability for personalized learning, tutoring, and feedback at scale.

Limitations

Our benchmark is written in English and focuses on common K–12 mathematics, so results may not transfer directly to other languages, curricula, or educational settings. MALRULELIB targets procedural malrules that can be implemented as executable transformations over symbolic expressions. This design is grounded in a review of 67 learning-science and mathematics education sources and yields broad coverage of prevalent K–12 procedural misconceptions (101 malrules and 498 templates across 22 categories), but it is not exhaustive. In particular, we do not aim to represent higher-level mathematics or student difficulties that are not well captured by a single procedure, such as conceptual misunderstandings, metacognitive errors, or strategy shifts.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Nira Almog and Bat-Sheva Ilany. 2012. [Absolute value inequalities: high school students’ solutions and misconceptions](#). *Educational Studies in Mathematics*, 81:347–364.
- John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tu-

- tors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207.
- Funda Aydın-Güç and Derya Aygün. 2021. Errors and misconceptions of eighth-grade students regarding operations with algebraic expressions. *International Online Journal of Education and Teaching (IOJET)*, 8(2):1106–1126. ERIC EJ1294052.
- Nina G. Bailey and Candice M. Quinn. 2023. Calculus ii students' understanding of the univalence requirement of function. In *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education (RUME)*. Research examining student reasoning about function univalence; NSF PURL: <https://par.nsf.gov/servlets/purl/10114085>.
- M. J. Behr, R. Lesh, T. Post, and E. Silver. 1983. Rational number, ratio, and proportion. In R. Lesh and M. Landau, editors, *Acquisition of mathematics concepts and processes*, pages 91–126. Academic Press.
- Merlyn J Behr, Ipke Wachsmuth, Thomas R Post, and Richard Lesh. 1984. Order and equivalence of rational numbers: A clinical teaching experiment. *Journal for Research in Mathematics Education*, 15(5):323–341.
- Samuel Kojo Biney, Clement Ali, and Nixon Adzifome. 2023. Errors and misconceptions in solving linear inequalities in one variable. *Journal of Advanced Sciences and Mathematics Education*, 3.
- John D Bransford, Ann L Brown, and Rodney R Cocking. 1999. *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington, DC.
- Ryan Britt and Melissa Weinrich. 2025. The price of precision: Significant figures and the student experience. *Journal of Chemical Education*, 102(11):4714–4724.
- George Brown and Robert J. Quinn. 2006. Algebra students' difficulty with fractions: An error analysis. *Australian Mathematics Teacher*, 62(4):28–40. ERIC EJ765838.
- John Seely Brown and Richard R Burton. 1978a. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2(2):155–192.
- John Seely Brown and Richard R. Burton. 1978b. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2(2):155–192.
- John Seely Brown and Richard R. Burton. 1978c. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2(2):155–192.
- John Seely Brown and Kurt VanLehn. 1980. Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, 4(4):379–426.
- L. Brown, R. Burton, and K. VanLehn. 1986. Identification and remediation of children's subtraction errors: A comparison of practical approaches. *Journal of School Psychology*, 24(2):163–180.
- Richard Cangelosi, Silvia Madrid, Sandra Cooper, Jo Olson, and Beverly Hartter. 2013a. The negative sign and exponential expressions: Unveiling students' persistent errors and misconceptions. *Journal of Mathematical Behavior*, 32(1):69–82.
- Richard Cangelosi, Silvia Madrid, Sandra Cooper, Jo Olson, and Beverly Hartter. 2013b. The negative sign and exponential expressions: Unveiling students' persistent errors and misconceptions. *The Journal of Mathematical Behavior*, 32(1):69–82.
- Richard Cangelosi, Silvia Madrid, Sandra Cooper, Jo Olson, and Beverly Hartter. 2013c. The negative sign and exponential expressions: Unveiling students' persistent errors and misconceptions. *The Journal of Mathematical Behavior*, 32(1):69–82.
- A. E. Castro Sotos, S. Vanhoof, W. Van den Noortgate, and P. Onghena. 2007. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 2(2):98–113.
- Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.
- Dirk De Bock, Wim Van Dooren, D. Janssens, L. Verschaffel, and J. Torbeys. 2002a. Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics*, 50:311–334.
- Dirk De Bock, Wim Van Dooren, Dirk Janssens, and Lieven Verschaffel. 2002b. Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics*, 50(3):311–334.
- Emrah Oguzhan Dincer and Aslihan Osmanoglu. 2018. Dealing with metric unit conversion: An examination of prospective science teachers' knowledge of and difficulties with conversion. *Science Education International*, 29(3):174–182. Published Sep 2018.
- Gunawardena Egodawatte Arachchige Don. 2011a. *Secondary School Students' Misconceptions in Algebra*. Phd dissertation, University of Toronto. Unpublished doctoral thesis.
- Gunawardena Egodawatte Arachchige Don. 2011b. *Secondary School Students' Misconceptions in Algebra*. Phd dissertation, University of Toronto. Unpublished doctoral thesis.
- Wim Van Dooren, Dirk De Bock, and Lieven Verschaffel. 2010. From addition to multiplication . . . and back: The development of students' additive and multiplicative reasoning skills. *Cognition and Instruction*, 28(3):360–381.
- J Al Easley and Russell E Zwoyer. 1975. Teaching by listening-toward a new day in math classes. *Contemporary Education*, 47(1):19.

- Joanne Eaves, Nina Attridge, and Camilla Gilmore. 2025. [Misconceptions of the order of operations and associativity use](#). *Learning and Instruction*, 97:102074.
- Karen C. Fuson, Sandra Smith, and Andrea Lo Cicero. 1997. [Understanding the structure of multi-digit addition and subtraction: The role of base-10 place value](#). *Journal for Research in Mathematics Education*, 28(2):131–162. Documents children’s tendency to write entire column sums without regrouping, violating base-10 place value principles.
- Peter L. Glidden. 2008. [Prospective elementary teachers’ understanding of order of operations](#). *School Science and Mathematics*, 108(4):130–136.
- Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The assistants ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497.
- James Hiebert and Diana Wearne. 1985. [A model of students’ decimal computation procedures](#). *Cognition and Instruction*, 2(3-4):175–205.
- A. M. Jarrah, Y. Wardat, and S. Gningue. 2022. [Misconception on addition and subtraction of fractions in seventh-grade middle school students](#). *Eurasia Journal of Mathematics, Science and Technology Education*, 18(6):em2115.
- Mihriban Hacısalihoğlu Karadeniz and Yasemin Çalışkan. 2023. [What are the misconceptions about scientific notation of very large and very small numbers?](#) *International Journal of Educational Studies in Mathematics*, 10(2):142–165.
- Robert Karplus, Stephen Pulos, and E.K. Stage. 1983. [Early adolescents’ proportional reasoning on ‘rate’ problems](#). *Educational Studies in Mathematics*, 14:219–233.
- Carolyn Kieran. 1981. [Concepts associated with the equality symbol](#). *Educational Studies in Mathematics*, 12:317–326.
- Igor’ Kontorovich. 2016. [The answer depends on your lecturer](#). *Research in Mathematics Education*, 18(3):283–298.
- Gaea Leinhardt, Orit Zaslavsky, and Mary Kay Stein. 1990. [Functions, graphs, and graphing: Tasks, learning and teaching](#). *Review of Educational Research*, 60(1):1–64.
- Naiming Liu, Shashank Sonkar, Zichao Wang, Simon Woodhead, and Richard G Baraniuk. 2023. [Novice learner and expert tutor: Evaluating math reasoning abilities of large language models with misconceptions](#). *arXiv preprint arXiv:2310.02439*.
- Li Lucy, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. [Evaluating language model math reasoning via grounding in educational curricula](#). *Preprint*, arXiv:2408.04226.
- Malcolm MacGregor and Kaye Stacey. 1997. [Students’ understanding of algebraic notation: 11–15](#). *Educational Studies in Mathematics*, 33:1–19.
- F. M. Machaba. 2016. [The concepts of area and perimeter: Insights and misconceptions of grade 10 learners](#). *Pythagoras*, 37(1):a304.
- Nancy K. Mack. 1995. [Confounding whole-number and fraction concepts when building on informal knowledge](#). *Journal for Research in Mathematics Education*, 26(5):422–441.
- Judah Paul Makonye and Josiah Fakude. 2016. [A study of errors and misconceptions in the learning of addition and subtraction of directed numbers in grade 8](#). *Sage Open*, 6(4):2158244016671375.
- Marilyn Matz. 1980. [Towards a computational theory of algebraic competence](#). *The Journal of Mathematical Behavior*, 3(1):93–166.
- Janice Mokros and Susan Jo Russell. 1995. [Children’s concepts of average and representativeness](#). *Journal for Research in Mathematics Education*, 26(1):20–39.
- K. J. Newton, C. Willard, and C. Teufel. 2014. [An examination of the ways that students with learning disabilities solve fraction computation problems](#). *The Elementary School Journal*, 115(1):1–21.
- Mbazima Amos Ngoveni. 2025. [Deconstructing and addressing factorizing errors and misconceptions in a tvet college: Mathematical insights and interventions](#). *The International Journal of Science, Mathematics and Technology Learning*, 32(2):23–47.
- Yiping Ni and Yifeng Zhou. 2005a. [Teaching and learning fractions and rational numbers: The origins and implications of whole-number bias](#). *Educational Psychologist*, 40(1):27–52.
- Yiping Ni and Yifeng Zhou. 2005b. [Teaching and learning fractions and rational numbers: The origins and implications of whole-number bias](#). *Educational Psychologist*, 40(1):27–52.
- Yujing Ni and Yong-Di Zhou. 2005c. [Teaching and learning fraction and rational numbers: The origins and implications of whole number bias](#). *Educational Psychologist*, 40(1):27–52.
- Michael C. Oehrtman, Marilyn P. Carlson, and Patrick W. Thompson. 2008a. [Foundational reasoning abilities that promote coherence in students’ understandings of function](#). In Marilyn P. Carlson and Chris Rasmussen, editors, *Making the Connection: Research and Practice in Undergraduate Mathematics*, pages 27–42. Mathematical Association of America, Washington, DC.

- Michael C. Oehrtman, Marilyn P. Carlson, and Patrick W. Thompson. 2008b. [Foundational reasoning abilities that promote coherence in students' understandings of function](#). In Marilyn P. Carlson and Chris Rasmussen, editors, *Making the Connection: Research and Practice in Undergraduate Mathematics*, pages 27–42. Mathematical Association of America, Washington, DC.
- John P. Papadouris, Vasileios Komis, and Konstantinos Lavidas. 2025. [Errors and misconceptions of secondary school students in absolute values: a systematic literature review](#). *Mathematics Education Research Journal*, 37:507–528.
- Melanie Parker and Gaea Leinhardt. 1995. [Percent: A privileged proportion](#). *Review of Educational Research*, 65(4):421–481.
- Demetra Pitta-Pantazi, Constantinos Christou, and Theodossios Zachariades. 2007a. [Secondary school students' levels of understanding in computing exponents](#). *The Journal of Mathematical Behavior*, 26(4):301–311.
- Demetra Pitta-Pantazi, Constantinos Christou, and Theodossios Zachariades. 2007b. [Secondary school students' levels of understanding in computing exponents](#). *The Journal of Mathematical Behavior*, 26(4):301–311.
- Yasseen Rabab'ah. 2025. [The most common conceptual errors in primary school geometry](#). *Educational Process: International Journal*, 18:e2025525. ERIC EJ1486282.
- Lauren Resnick, Pearla Nesher, François Léonard, Maria Magone, Susan Omanson, and Irit Peled. 1989a. [Conceptual bases of arithmetic errors: The case of decimal fractions](#). *Journal for Research in Mathematics Education*, 20.
- Lauren B Resnick, Pearla Nesher, François Leonard, Maria Magone, Susan Omanson, and Irit Peled. 1989b. [Conceptual bases of arithmetic errors: The case of decimal fractions](#). *Journal for Research in Mathematics Education*, 20(1):8–27.
- Robert S Siegler, Greg J Duncan, Pamela E Davis-Kean, Kathryn Duckworth, Amy Claessens, Mimi Engel, Maria Ines Susperreguy, and Meichu Chen. 2012. [Early predictors of high school mathematics achievement](#). *Psychological Science*, 23(7):691–697.
- D. Sleeman. 1984. [An attempt to understand students' understanding of basic algebra](#). *Cognitive Science*, 8(4):387–412.
- Shashank Sonkar and Richard G Baraniuk. 2023. [Deduction under perturbed evidence: Probing student simulation \(knowledge tracing\) capabilities of large language models](#). In *LLM@ AIED*, pages 26–33.
- Shashank Sonkar, Naiming Liu, Xinghe Chen, and Richard Baraniuk. 2025. [Turing-like test for personalized educational ai](#). In *International Conference on Artificial Intelligence in Education*, pages 405–412. Springer.
- Shashank Sonkar, Naiming Liu, MyCo Le, and Richard Baraniuk. 2024a. [Malalogoqa: Pedagogical evaluation of counterfactual reasoning in large language models and implications for ai in education](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15554–15567.
- Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. [Class: A design framework for building intelligent tutoring systems based on learning science principles](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961.
- Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024b. [Pedagogical alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650.
- Shashank Sonkar, Andrew E Waters, Andrew S Lan, Phillip J Grimaldi, and Richard G Baraniuk. 2020. [qdkit: Question-centric deep knowledge tracing](#). *arXiv preprint arXiv:2005.12442*.
- Stamatia Stafylidou and Stella Vosniadou. 2004. [The development of students' understanding of the numerical value of fractions](#). *Learning and Instruction*, 14(5):503–518. The Conceptual Change Approach to Mathematics Learning and Teaching.
- Vicki Steinle and Kaye Stacey. 1998a. [The incidence of misconceptions of decimal notation amongst students in grades 5 to 10](#). *Teaching Mathematics in New Times*, 2.
- Vicki Steinle and Kaye Stacey. 1998b. [The incidence of misconceptions of decimal notation amongst students in grades 5 to 10](#). *Teaching Mathematics in New Times*, 2.
- Sheryl L Stump. 2001. [Developing preservice teachers' pedagogical content knowledge of slope](#). *The Journal of Mathematical Behavior*, 20(2):207–227.
- Sanem Tabak. 2019. [6th, 7th and 8th grade students' misconceptions about the order of operations](#). *Başlık*, volume-5-2019(volume5-issue3.html):363–373.
- G. Tan Sisman and M. Aksu. 2016a. [A study on sixth grade students' misconceptions and errors in spatial measurement: Length, area, and volume](#). *International Journal of Science and Mathematics Education*, 14:1293–1319.
- G. Tan Sisman and M. Aksu. 2016b. [A study on sixth grade students' misconceptions and errors in spatial measurement: Length, area, and volume](#). *International Journal of Science and Mathematics Education*, 14:1293–1319.

- Jane Tendere and Lillias H. N. Mutambara. 2020. [An analysis of errors and misconceptions in the study of quadratic equations](#). *European Journal of Mathematics and Science Education*, volume-1-2020(volume-1-issue-2-december-2020):81–90.
- Dina Tirosh. 2000a. [Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions](#). *Journal for Research in Mathematics Education*, 31(1):5–25.
- Dina Tirosh. 2000b. [Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions](#). *Journal for Research in Mathematics Education*, 31(1):5–25.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Fadime Ulusoy. 2019a. Serious obstacles hindering middle school students' understanding of integer exponents. *International Journal of Research in Education and Science*, 5(1):52–69. ERIC EJ1198050.
- Fadime Ulusoy. 2019b. Serious obstacles hindering middle school students' understanding of integer exponents. *International Journal of Research in Education and Science*, 5(1):52–69. ERIC EJ1198050.
- Fadime Ulusoy. 2019c. [Serious obstacles hindering middle school students' understanding of integer exponents](#). *International Journal of Research in Education and Science*, 5(1):60–68.
- Kristof Van Hoof, Melissa DeWolf, Lieven Verschaffel, and Wim Van Dooren. 2021. [Examining the relation between whole numbers and fractions: Whole-number bias in fraction operations](#). *Learning and Instruction*, 76:101526.
- Maria Varelas and Joe Becker. 1997. [Children's developing understanding of place value: Semiotic aspects](#). *Cognition and Instruction*, 15(2):265–286. Documents children's difficulties differentiating between face value and complete value of digits, providing evidence of place-value misconceptions.
- Joelle Vlassis. 2004a. Making sense of the minus sign or becoming flexible in 'negativity'. *Learning and Instruction*, 14(5):469–484.
- Joëlle Vlassis. 2004b. [Making sense of the minus sign or becoming flexible in 'negativity'](#). *Learning and Instruction*, 14(5):469–484. The Conceptual Change Approach to Mathematics Learning and Teaching.
- Xiong Wang. 2015. [The literature review of algebra learning: Focusing on the contributions to students' difficulties](#). *Creative Education*, 6(20):2274–2285.
- Eamon Worden, Cristina Heffernan, Neil Heffernan, and Shashank Sonkar. 2026. Foundationalassist: An educational dataset for foundational knowledge tracing and pedagogical grounding of llms. *arXiv preprint arXiv:2602.00070*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Erdogan Mehmet Özkan. 2011. [Misconceptions in radicals in high school mathematics](#). *Procedia - Social and Behavioral Sciences*, 15:120–127. 3rd World Conference on Educational Sciences - 2011.

Appendix

A Additional Malrule Examples

Table 7 expands on Table 2 with additional malrule examples across 10 categories, each showing two template variations demonstrating cross-template generalization.

B Malrules and Their Sources

Tables 8–11 present the source and description for each malrule in this study.

C Full Results

C.1 Per-Malrule Breakdown

Tables 12–14 present performance breakdown for all 101 malrules, sorted by accuracy.

C.2 Template Listing

Tables 17–19 list all templates for each malrule, organized by category.

D Template Listing

This section provides comprehensive analysis of all 498 templates across 101 malrules. The template design reflects intentional pedagogical choices grounded in learning science research, enabling rigorous cross-template generalization testing.

D.1 Template Statistics Overview

- **Total templates:** 498 across 101 malrules (4.9 average per malrule)
- **Context domains:** 10 domains (63.3% abstract, 36.7% contextualized)—grounded in *transfer of learning* research
- **Scaffold levels:** Basic (18.5%), Variant (50.8%), Context (6.2%), Word Problem (24.5%)—mirrors classroom instruction progression

D.2 Context Domain Examples

Templates embed problems in diverse real-world contexts (Table 15). This design tests *transfer of learning*—research shows students often fail to apply knowledge across contexts (Bransford et al., 1999). A student may correctly add fractions abstractly but fail when the same problem appears as pizza slices.

D.3 Scaffolded Complexity

Each malrule follows a pedagogical progression mirroring classroom instruction (Table 16). This scaffolding enables assessment across difficulty gradients—word problems in particular test whether misconceptions persist despite additional cognitive load from reading comprehension.

D.4 Templates by Malrule

Tables 17–20 list all templates for each malrule, organized by category.

E Framework Details

This section provides additional details on the MALRULELIB framework template and classification coverage.

E.1 Template Coverage

Table 21 provides detailed template coverage statistics for each mathematical category.

E.2 Classification Details

Table 22 shows the distribution of malrules across NCTM strands.

Category	Malrule	Problem	Student's Work
Statistics	Mode must exist	T1: Identify the mode for the following data set: [15, 95, 36, 15, 4, 82] T2: Dataset A: [2, 8, 41, 41], Dataset B: [68, 65, 52, 66, 91, 58]. Which dataset has a mode?	Frequencies: 4:1, 15:2, 36:1, 82:1, 95:1 → 95 Dataset A: Mode = 41(correct), Dataset B: Mode = 91(picked biggest)
Algebra	$x + a = b \rightarrow x = b + a$	T1: Solve for x: $4x + 9 - 7 = 34$ T2: A phone plan costs \$29 monthly plus \$5 per GB of data. If the bill is \$114, how many GB (x) were used?	$4x + 9 - 7 = 34 \rightarrow 4x = 34 + 9 - 7;$ $\rightarrow 4x = 36 \rightarrow x = 9$ $5x + 29 = 114 \rightarrow 5x = 114 + 29$ $\rightarrow 5x = 143 \rightarrow x = 28$
Scientific Notation	Count all zeros for exponents	T1: Write 0.010500 in scientific notation. T2: Is 0.00002050 equal to 2.05×10^{-7} ?	Zeros count: 5, Coefficient: 1.05 → Exponent: -5, 1.05×10^{-5} Zeros count: 7, Coefficient: 2.05, Exponent: -7, Answer: Yes
Absolute Value	$ a+b = a + b $	T1: Evaluate $ x - 2 $ when $x = 10$ T2: A drone is at position x meters, and moves 3 meters to the right. The distance from the origin is $ x + 3 $. If $x = -3$, what is the distance from the origin?	$ x - 2 = x - 2 = 10 - 2 = 8$ $ x + 3 = x + 3 = 3 + 3 = 6$ meters
Decimals	More digits → larger value	T1: Which is longer: 0.5 kilometers or 0.479 kilometers? T2: Maria has \$0.61 and Tom has \$0.214. Who has more money?	0.5: 1 places, 0.479: 3 places, $3 > 1$, 0.479 0.61: 2 places, 0.214: 3 places, $3 > 2$, Tom
Exponents	$(a + b)^n = a^n + b^n$	T1: Evaluate: $(1 + 3 + 2)^2$ T2: A server processes 2 GB in phase 1 and 4 GB in phase 2 (total: 6 GB). This data volume is replicated across 2 redundant systems with exponential scaling. Calculate $(2 + 4)^2$.	$(1 + 3 + 2)^2 = 1^2 + 3^2 + 2^2 = 1 + 9 + 4 = 14$ $(2 + 4)^2 = 2^2 + 4^2 = 4 + 16 = 20$
Factoring	$a^2 + b^2 = (a + b)^2$	T1: Factor: $x^2 + 36$ T2: Factor: $3^2x^2 + 5^2y^2$	$(x + 6)^2 = x^2 + 36$, $(x + 6)^2$ $(3x + 5y)^2 = 3^2x^2 + 5^2y^2$, $(3x + 5y)^2$
Fractions	Add numerators together and denominators together	T1: What is $\frac{5}{3} + \frac{7}{4}$? T2: Sarah ate $\frac{1}{4}$ of a pizza and John ate $\frac{1}{3}$ of the same pizza. What fraction of the pizza did they eat together?	$\frac{5}{3} + \frac{7}{4}$, $5 + 7 = 12$, $3 + 4 = 7 \rightarrow \frac{12}{7}$ $\frac{1}{4} + \frac{1}{3}$, $1 + 1 = 2$, $4 + 3 = 7 \rightarrow \frac{2}{7}$
Geometry	Surface area = $l \times w \times h$	T1: Find the surface area of a rectangular prism with length 4.1 cm, width 5.4 cm, and height 3.0 cm. T2: A storage container measures 4 feet long, 8 feet wide, and 8 feet tall. What is the total surface area that needs to be painted?	$A = l \times w \times h = 4.1 \times 5.4 \times 3.0 = 66.42$ $A = l \times w \times h = 4 \times 8 \times 8 = 256$
Linear Equations	Slope = $\frac{\Delta x}{\Delta y}$	T1: Find the slope of the line passing through points (-9, 2) and (1, 8). T2: After 15 hours, a vehicle has gone 18 miles. After 19 hours, it has gone 5 miles. Calculate the speed.	$\Delta x = 10$, $\Delta y = 6$, Slope = $\frac{10}{6} = \frac{5}{3}$ Point 1 = (15, 18), Point 2 = (19, 5), $\Delta x = 4$, $\Delta y = -13$, Slope = $\frac{4}{-13} = -\frac{4}{13}$

Table 7: More Examples of malrules from MALRULELIB, showing two templates per misconception.

Malrule (Source)	Description
absolute_value_distributes (Papadouris et al., 2025)	Treating $ a + b $ as $ a + b $
absolute_value_makes_positive (Papadouris et al., 2025)	Students hold the mental picture that 'absolute values make negative signs positive' and apply this symbol-based process rather than understanding that absolute value represents distance from zero, treats $ x = -a$ as $x = a$
inequality_direction_confusion (Almog and Ilany, 2012)	Students confuse the solution patterns for absolute value inequalities, applying the 'less than' pattern (three-part-inequality) to 'greater than' problems and vice versa.
cancel_across_equals (Sleeman, 1984)	Students incorrectly 'cancel' matching variable terms from both sides of an equation as if they were canceling factors in a fraction, rather than properly subtracting the terms from both sides.
change_side_change_sign (Kieran, 1981)	Students incorrectly believe that when moving a term from one side of an equation to the other, no sign need to change, i.e., $x + a = b \rightarrow x = b + a$
distribute_over_non_distributive (Sleeman, 1984)	Students incorrectly extend the distributive property to operations that do not distribute, such as exponentiation and square roots.
divide_one_term_only (Wang, 2015)	Students incorrectly apply division to only one term (typically the variable term) instead of distributing the operation to all terms on that side of the equation.
forget_negative_division (Biney et al., 2023)	Students neglect the crucial rule that multiplying or dividing an inequality by a negative coefficient reverses the direction of the inequality sign.
variable_letter_has_value (MacGregor and Stacey, 1997)	Students believe that algebraic letters have inherent fixed values, often based on the letter's position in the alphabet (e.g., $x = 10$, $a = 1$) or other associations.
ignore_decimal_point (Resnick et al., 1989a)	Students ignore the decimal point and treat decimal numbers as whole numbers, applying whole number arithmetic procedures without regard to place value.
longer_is_larger (Steinle and Stacey, 1998a)	When comparing decimal numbers, students incorrectly believe that more decimal digits means a larger number.
right_align_decimals (Hiebert and Wearne, 1985)	Students align decimal numbers by the rightmost digit instead of by the decimal point when performing addition or subtraction.
shorter_is_larger (Steinle and Stacey, 1998b)	Students incorrectly believe that decimals with fewer decimal places are larger.
whole_number_thinking (Resnick et al., 1989a)	Students incorrectly treat decimals as if they demonstrate the properties of whole numbers. For example, $3.7 + 2.5$ treated as $37 + 25 = 62$, $0.45 > 0.8$ because $45 > 8$
add_exponents_for_power_of_power (Pitta-Pantazi et al., 2007a)	Student adds exponents instead of multiplying them when computing power of a power.
distribute_exponent_over_addition (Don, 2011a)	Student incorrectly distributes exponents over addition: $(a + b)^n = a^n + b^n$.
forget_exponent_on_coefficient (Cangelosi et al., 2013a)	Student forgets to raise coefficient to the outer power: $(cx^m)^n = c \cdot x^{mn}$.
multiply_base_by_exponent (Ulusoy, 2019a)	Student treats exponentiation as multiplication: $a^b = a \times b$.
multiply_exponents_when_multiplying_powers (Pitta-Pantazi et al., 2007b)	Student multiplies exponents when multiplying powers with the same base: $x^m \times x^n = x^{m \times n}$.
negative_exponent_makes_negative (Cangelosi et al., 2013b)	Student thinks negative exponent makes the result negative: $x^{-n} = -x^n$.
zero_exponent_equals_zero (Ulusoy, 2019b)	Student thinks $a^0 = 0$ instead of $a^0 = 1$.
incomplete_factoring (Ngoveni, 2025)	Stop after first factoring step when the result is not fully factored

Table 8: Source and description for each malrule (Part 1 of 4).

Malrule (Source)	Description
sign_errors_in_factoring (Tendere and Mutambara, 2020)	Students make sign errors when factoring quadratics, especially in decomposition method. Such as $a^2 - b^2 = (a - b)(a + b)$, perfect square, and factoring constants sign.
sum_of_squares_factors (Don, 2011b)	Student thinks $a^2 + b^2 = (a + b)^2$
add_numerators_add_denominators (Jarrah et al., 2022)	When adding fractions with different denominators, students incorrectly add numerators together and denominators together, treating the operation as component-wise addition.
common_denominator_numerator (Mack, 1995)	This is a variant of the add-numerators-add-denominators error that shows partial understanding - students know they need a common denominator but incorrectly believe adding denominators produces one.
denominator_comparison_error (Stafylidou and Vosniadou, 2004)	Students incorrectly compare fractions by focusing on the denominator value, believing that a larger denominator means a larger fraction.
ignore_denominators (Van Hoof et al., 2021)	Students operate only on numerators, completely ignoring denominators, treating fraction operations as whole-number arithmetic on the 'top numbers' only.
keep_common_denominator_for_multiplication (Newton et al., 2014)	Students incorrectly keep the common denominator when multiplying fractions with like denominators: $(a/b) \times (c/b) = (a \times c)/b$ instead of $(a \times c)/(b \times b)$.
multiply_across_for_division (Tirosh, 2000a)	Student treats fraction division like multiplication, forgetting to invert (flip) the second fraction before multiplying.
natural_number_bias_numerator_only (Ni and Zhou, 2005a)	Students incorrectly compare fractions by comparing their numerators only, ignoring the denominators entirely. They believe that a larger numerator means a larger fraction.
subtract_across (Brown and Quinn, 2006)	When subtracting fractions with different denominators, students incorrectly subtract the numerators AND subtract the denominators: $\frac{a}{b} - \frac{c}{d} = (a - c)/(b - d)$.
function_distributive_property (De Bock et al., 2002a)	Student incorrectly applies the additive property of linear functions to nonlinear functions: $f(x + a) = f(x) + f(a)$.
function_notation_is_multiplication (Oehrtman et al., 2008a)	Student interprets $f(x)$ as meaning $f \times x$ (multiplication) rather than function notation.
same_input_different_outputs_ok (Bailey and Quinn, 2023)	Students fail to understand the univalence requirement: that functions must map each input to exactly ONE output.
scalar_multiplication_inside_or_outside_same (Oehrtman et al., 2008b)	Student incorrectly applies the multiplicative (homogeneous) property of linear functions to nonlinear functions: $f(cx) = c \cdot f(x)$.
count_net_perimeter_as_surface_area (Tan Sisman and Aksu, 2016a)	Measure net perimeter instead of calculating surface area
same_area_same_perimeter (Machaba, 2016)	Students think same area means same perimeter.
same_perimeter_same_area (Machaba, 2016)	Students think same perimeter means same area.
volume_formula_for_surface_area (Tan Sisman and Aksu, 2016b)	Students use volume formula $V = lwh$ when asked for surface area.
ignore_coordinate_signs (Rabab'ah, 2025)	Student ignores or misuses the signs of coordinates, treating negative values as positive (absolute values) when plotting points or identifying quadrants.
confuse_slope_and_intercept_roles (Leinhardt et al., 1990)	Student swaps the roles of slope (m) and y-intercept (b) in $y = mx + b$, either writing equations with parameters reversed or identifying the constant term as slope and the coefficient as y-intercept.
slope_direction_confusion (Leinhardt et al., 1990)	Student confuses positive and negative slope directions, incorrectly stating that negative slopes represent increasing functions or positive slopes represent decreasing functions.
slope_is_delta_x_over_delta_y (Stump, 2001)	Student inverts the slope formula, calculating $\frac{\Delta x}{\Delta y}$ instead of $\frac{\Delta y}{\Delta x}$.

Table 9: Source and description for each malrule (Part 2 of 4).

Malrule (Source)	Description
inverted_conversion_factor (Dincer and Osmanoglu, 2018)	Students set up the conversion factor incorrectly by placing units in the wrong position (numerator vs. denominator), resulting in multiplication when division is required or vice versa.
wrong_conversion_factor (Dincer and Osmanoglu, 2018)	Students use incorrect conversion factors when converting between units, often substituting convenient round numbers (10, 100) or powers of ten for the actual conversion factors.
alignment_error_in_multi_digit (Brown and Burton, 1978b)	When multiplying by the tens digit, students do not shift the partial product one place to the left, treating the tens digit as if it were in the ones place.
divide_larger_by_smaller_always (Tirosh, 2000b)	A persistent misconception where students believe division must always involve a larger dividend divided by a smaller divisor. When presented with problems where the dividend is smaller than the divisor (e.g., $4 \div 6$), students either reverse the operands or claim the problem is impossible.
division_makes_smaller (Tirosh, 2000b)	A widespread misconception where students believe division always produces a result smaller than the dividend.
forget_to_add_carried_number (Brown and Burton, 1978b)	A systematic procedural bug where students correctly multiply each digit but forget to add the carried (regrouped) value to the next place value.
multiplication_makes_bigger (Tirosh, 2000b)	A widespread misconception where students believe multiplication always produces a result larger than both factors.
larger_absolute_value_always_wins (Makonye and Fakude, 2016)	Students incorrectly determine the sign of the result by always using the sign of the number with the larger absolute value, regardless of the operation.
multiplication_rule_for_addition (Makonye and Fakude, 2016)	Students incorrectly apply multiplication sign rules (like 'two negatives make a positive') to addition and subtraction operations.
negative_swaps_operation (Vlassis, 2004b)	Students confuse the negative sign's role as a number property with the operation being performed, leading to incorrect operation swapping.
negative_times_negative_negative (Cangelosi et al., 2013c)	Students incorrectly believe that multiplying two negative numbers gives a negative result, reversing the correct sign rule.
two_negatives_always_positive (Makonye and Fakude, 2016)	Students incorrectly apply the multiplication rule 'two negatives make a positive' to subtraction and addition operations.
addition_before_subtraction-always (Eaves et al., 2025)	Students incorrectly believe that addition must ALWAYS be performed before subtraction.
ignore_parentheses (Aydın-Güç and Aygün, 2021)	Students ignore parentheses and evaluate left-to-right or apply PEMDAS without respecting grouping symbols.
multiplication_before_division_always (Eaves et al., 2025)	Students incorrectly believe that multiplication must ALWAYS be performed before division.
pemdas_strictly_sequential (Glidden, 2008)	Students treat PEMDAS as a strict six-step sequence rather than understanding it as four priority levels.
strict_left_to_right (Tabak, 2019)	Students evaluate arithmetic expressions strictly from left to right, ignoring operator precedence rules (PEMDAS/BODMAS).
add_percentages_directly (Dooren et al., 2010)	Students incorrectly apply additive reasoning to percentage problems that require multiplicative thinking. They add or subtract percentages directly without recognizing that each percentage applies to a different base value.
percent_equals_decimal (Parker and Leinhardt, 1995)	Students treat percent, decimal, and whole number notations as interchangeable, failing to recognize that percent means 'per hundred' and requires division by 100 to convert to decimal form.
percentage_as_index (Parker and Leinhardt, 1995)	Students treat the percentage number as an absolute value or index, ignoring the base/whole that the percentage applies to.
reverse_percentage_error (Parker and Leinhardt, 1995)	Students incorrectly believe that percentage relationships are symmetric or reversible. They assume that if a value increased by X%, it can return to the original by decreasing X%, or that if A is X% of B, then B is X% of A.
add_under_common_root (De Bock et al., 2002b)	Students combine radicals under common root, i.e., students believes that $\sqrt{a} - \sqrt{b} = \sqrt{a - b}$.

Table 10: Source and description for each malrule (Part 3 of 4).

Malrule (Source)	Description
distribute_square_root_over_addition (De Bock et al., 2002b)	Students believe that square root distributes over addition, $\sqrt{}$ applies to each term separately, i.e. $\sqrt{a^2 + b^2} = \sqrt{a^2} + \sqrt{b^2}$
negative_outside_same_as_inside (Özkan, 2011)	Students confuse $-\sqrt{n}$ with $\sqrt{-n}$, failing to recognize that the domain of the radical function is restricted to nonnegative real numbers.
square_root_equals_plus_minus (Kontorovich, 2016)	Students incorrectly believe that the radical symbol \sqrt{n} yields both positive and negative values ($\pm\sqrt{n}$), confusing the principal square root with the solutions to equations of the form $x^2 = n$.
additive_instead_of_multiplicative (Karplus et al., 1983)	Student applies additive reasoning instead of multiplicative reasoning when working with proportions. The student computes the difference between values and adds this constant to find missing values, rather than using the multiplicative scale factor.
each_fraction_digit_is_ratio (Ni and Zhou, 2005b)	Student treats the numerator and denominator as independent whole numbers rather than as components of a single rational number.
ratio_as_division_only (Behr et al., 1983)	Student interprets ratio solely as division (a quotient), failing to understand that a ratio represents a multiplicative comparison between two quantities.
swap_ratios_or_units (Behr et al., 1983)	Student incorrectly sets up proportion equations by placing values in wrong positions, inverting the relationship between quantities.
decimal_places_same_as_sig_figs (Britt and Weinrich, 2025)	Students confuse decimal places with significant figures. When asked to round to N significant figures, they instead round to N decimal places.
add_coefficients_when_multiplying (Ulusoy, 2019c)	When multiplying numbers in scientific notation, students correctly add the exponents but incorrectly add the coefficients instead of multiplying them.
count_all_zeros_for_exponent (Karadeniz and Çalıřkan, 2023)	Students count ALL zeros in a number when determining the exponent for scientific notation, without considering their placement.
wrong_exponent_sign (Ulusoy, 2019c)	Students use the wrong sign for the exponent when converting numbers to scientific notation.
ignore_outliers_effect (Castro Sotos et al., 2007)	Student tend to ignore outliers and assume mean is always representative
mean_without_understanding (Castro Sotos et al., 2007)	Student always use mean as the primary measure regardless the case.
mode_must_exist (Mokros and Russell, 1995)	Students force a mode to exist by picking largest/middle/smallest value in the data.
always_borrow_left (Brown and Burton, 1978c)	Students always borrow from the left column in subtraction, even when the top digit is greater than or equal to the bottom digit and borrowing is not needed.
borrow_from_bottom (Brown et al., 1986)	When borrowing is needed, students incorrectly decrement the subtrahend (bottom number) instead of the minuend (top number).
borrow_no_decrement (Brown and Burton, 1978c)	When borrowing is required, students correctly add 10 to the current column's digit but forget to decrement the digit in the column they borrowed from.
decompose_by_place_value_label (Varelas and Becker, 1997)	When interpreting place value decompositions, the student treats the numeric labels as face values to be concatenated rather than as multiplicative values to be added.
diff_0_n_equals_n (Brown and Burton, 1978c)	When the minuend digit is 0 and the subtrahend digit is N (where $N > 0$), the student writes N as the result instead of borrowing.
no_column_limit (Fuson et al., 1997)	When adding multi-digit numbers, students write the entire column sum directly in that column without regrouping, failing to understand the base-10 constraint that each place value position can only hold a single digit (0-9).
smaller_from_larger (Brown and Burton, 1978c)	A systematic procedural bug where the student always subtracts the smaller digit from the larger digit in each column, regardless of position (minuend or subtrahend).
stops_borrow_at_zero (Brown and VanLehn, 1980)	A systematic procedural bug where the student stops the borrowing (regrouping) process entirely when encountering a zero in the column to the left, rather than cascading the borrow further left to find a non-zero digit.

Table 11: Source and description for each malrule (Part 4 of 4).

Malrule	Total	MRA	Accuracy
subtraction.borrow_from_bottom	2700	132	4.89
algebra.variable_letter_has_value	2700	169	6.26
fractions.multiply_rule_for_addition	180	22	12.22
ratios_proportions.each_fraction_digit_is_ratio	2700	331	12.26
statistics.mean_without_understanding	3600	546	15.17
subtraction.carry_ones_digit_instead_of_tens	2700	544	20.15
absolute_value.absolute_value_makes_positive	2700	551	20.41
negative_numbers.negative_times_negative_negative	3240	750	23.15
scientific_notation.wrong_exponent_sign	2700	632	23.41
measurement.wrong_conversion_factor	180	43	23.89
radicals.add_under_common_root	3420	862	25.20
scientific_notation.count_all_zeros_for_exponent	2700	696	25.78
fractions.natural_number_bias_numerator_only	3060	836	27.32
factoring.sign_errors_in_factoring	180	50	27.78
negative_numbers.two_negatives_always_positive	2700	750	27.78
graphing.reverse_coordinate_order	180	51	28.33
graphing.ignore_coordinate_signs	2700	778	28.81
algebra.divide_one_term_only	180	52	28.89
subtraction.smaller_from_larger	4140	1234	29.81
subtraction.no_column_limit	2700	809	29.96
algebra.change_side_change_sign	180	55	30.56
order_of_operations.addition_before_subtraction_always	2700	829	30.70
order_of_operations.ignore_parentheses	2700	836	30.96
scientific_notation.add_coefficients_when_multiplying	2700	836	30.96
geometry.count_net_perimeter_as_surface_area	2700	839	31.07
subtraction.stops_borrow_at_zero	180	58	32.22
factoring.sum_of_squares_factors	2700	899	33.30
subtraction.skip_equal	2700	902	33.41
multiplication_division.forget_to_add_carried_number	3600	1277	35.47
negative_numbers.multiplication_rule_for_addition	2700	978	36.22
subtraction.always_borrow_left	180	67	37.22
algebra.cancel_across_equals	2700	1007	37.30
order_of_operations.multiplication_before_division_always	2700	1016	37.63
fractions.subtract_across	3060	1200	39.22

Table 12: Per-malrule performance breakdown (Part 1 of 3).

Malrule	Total	MRA	Accuracy
subtraction.borrow_no_decrement	180	71	39.44
fractions.keep_common_denominator_for_multiplication	3060	1243	40.62
factoring.negative_one_factor_forgotten	2880	1189	41.28
exponents.add_exponents_for_power_of_power	2880	1245	43.23
negative_numbers.negative_swaps_operation	2700	1177	43.59
multiplication_division.alignment_error_in_multi_digit	3600	1574	43.72
absolute_value.absolute_value_distributes	2700	1188	44.00
ratios_proportions.additive_instead_of_multiplicative	2880	1306	45.35
fractions.ignore_denominators	180	83	46.11
algebra.distribute_over_non_distributive	2700	1254	46.44
fractions.multiply_across_for_division	3240	1527	47.13
statistics.ignore_outliers_effect	2880	1388	48.19
algebra.forget_negative_division	180	87	48.33
ratios_proportions.swap_ratios_or_units	180	87	48.33
ratios_proportions.ratio_as_division_only	2880	1398	48.54
scientific_notation.ignore_different_powers_of_ten	2700	1327	49.15
measurement.inverted_conversion_factor	180	90	50.00
word_problems.include_all_numbers_given	3960	1997	50.43
subtraction.diff_0_n_equals_n	2700	1376	50.96
decimals.right_align_decimals	2880	1500	52.08
factoring.incomplete_factoring	2700	1415	52.41
linear_equations.slope_is_delta_x_over_delta_y	2700	1421	52.63
functions.function_notation_is_multiplication	180	95	52.78
decimals.longer_is_larger	2880	1540	53.47
order_of_operations.pemdas_strictly_sequential	2700	1450	53.70
percentages.percentage_as_index	2700	1465	54.26
absolute_value.inequality_direction_confusion	2160	1178	54.54
order_of_operations.strict_left_to_right	2700	1473	54.56
linear_equations.confuse_slope_and_intercept_roles	2700	1508	55.85
radicals.square_root_equals_plus_minus	3060	1743	56.96
negative_numbers.larger_absolute_value_always_wins	180	104	57.78
decimals.whole_number_thinking	2880	1673	58.09
decimals.ignore_decimal_point	2880	1718	59.65
radicals.square_root_is_divide_by_two	3060	1842	60.20

Table 13: Per-malrule performance breakdown (Part 2 of 3).

Malrule	Total	MRA	Accuracy
rounding.leading_zeros_are_significant	3600	2182	60.61
exponents.multiply_exponents_when_multiplying_powers	2880	1782	61.88
subtraction.subtract_smaller_from_larger_each_column	180	112	62.22
fractions.denominator_comparison_error	3060	1905	62.25
linear_equations.slope_direction_confusion	2880	1826	63.40
multiplication_division.multiplication_makes_bigger	3600	2299	63.86
factoring.forget_gcf_first	180	115	63.89
percentages.percent_equals_decimal	2700	1766	65.41
exponents.forget_exponent_on_coefficient	2880	1914	66.46
rounding.trailing_zeros_always_significant	3600	2401	66.69
fractions.common_denominator_numerator	2880	1943	67.47
multiplication_division.divide_larger_by_smaller_always	3600	2471	68.64
functions.scalar_multiplication_inside_or_outside_same	2700	1860	68.89
percentages.add_percentages_directly	180	124	68.89
exponents.distribute_exponent_over_addition	180	133	73.89
fractions.add_numerators_denominators	180	134	74.44
statistics.mode_must_exist	2628	1985	75.53
geometry.same_area_same_perimeter	180	136	75.56
multiplication_division.division_makes_smaller	3600	2778	77.17
decimals.shorter_is_larger	2880	2262	78.54
percentages.reverse_percentage_error	2700	2137	79.15
geometry.volume_formula_for_surface_area	2700	2168	80.30
linear_equations.y_intercept_always_positive	2700	2170	80.37
functions.function_distributive_property	180	146	81.11
radicals.distribute_square_root_over_addition	3060	2547	83.24
geometry.same_perimeter_same_area	180	150	83.33
subtraction.decompose_by_place_value_label	2700	2255	83.52
exponents.negative_exponent_makes_negative	2880	2415	83.85
rounding.decimal_places_same_as_sig_figs	180	153	85.00
exponents.multiply_base_by_exponent	2340	2022	86.41
radicals.negative_outside_same_as_inside	3060	2691	87.94
exponents.zero_exponent_equals_zero	2880	2714	94.24
functions.same_input_different_outputs_ok	2700	2628	97.33

Table 14: Per-malrule performance breakdown (Part 3 of 3).

Domain	Keywords	Example
Abstract (63%)	Pure math notation	“Calculate: $\frac{1}{2} + \frac{1}{3}$ ”
Measurement (11%)	Area, meters, feet	“A rectangle has length 2.1m and width 5.4m. What is its area?”
Money (9%)	Cost, price, dollar	“Sarah has \$0.5 and Tom has \$0.479. Who has more?”
Time/Distance (4%)	Speed, hours, miles	“A car travels at 5.4 km/hr for 4.3 hours. Total distance?”
Science (3%)	Bacteria, wavelength	“A wavelength of 1.13 μm is multiplied by 3.41”
Sports (3%)	Points, scores, team	“Team A has 4x points. Team B has 1x points...”
Food (3%)	Pizza, recipe, cake	“Sara ate $\frac{1}{4}$ of a pizza...”
Temperature (2%)	Degrees, heating	“The temperature was x degrees. It rose by 5...”
Sharing (1%)	Divided among	“If 12 cookies are shared among 4 friends...”
Elevation (1%)	Submarine, depth	“A submarine at -20m descends 15m more...”

Table 15: Context domain distribution across templates.

Level	Characteristics
Basic (18.5%)	Core mathematical formulation with simple values: “Calculate: $\frac{1}{2} + \frac{1}{3}$ ”
Variant (50.8%)	Structural variations—larger numbers, multiple operands, negative values, edge cases: “Calculate: $\frac{3}{4} + \frac{5}{6} + \frac{7}{8}$ ”
Context (6.2%)	Real-world scenario with units: “A rope is 41.24m long and another is 2.5m. Total length?”
Word Problem (24.5%)	Full story problem requiring comprehension: “Maria earned \$11.50 on Monday, \$8.75 on Tuesday, and spent \$4.30. How much does she have?”

Table 16: Scaffold level distribution across templates.

Category	Malrule	Templates
absolute_value	absolute_value_distributes	basic_addition, basic_subtraction, negative_result_inside, multiplication_inside, word_problem
	absolute_value_makes_positive	basic_equation, expression_inside, inequality, compound_expression, word_problem
algebra	inequality_direction_confusion	less_than, greater_than
	cancel_across_equals	basic_equation, word_problem_context, both_sides_constant, three_term_equation, comparison_problem
	change_side_change_sign	default
	distribute_over_non_distributive	basic_square_binomial, basic_sqrt_sum, square_binomial_subtraction, sqrt_difference, word_problem
decimals	divide_one_term_only	default
	forget_negative_division	default
	variable_letter_has_value	basic_addition, basic_subtraction, basic_multiplication, two_step_equation, word_problem
	ignore_decimal_point	basic_multiplication, division, scientific_context, money_context, measurement_context, word_problem_context
	longer_is_larger	basic_comparison, ordering, money_context, measurement_context, number_line, word_problem_context
	right_align_decimals	basic_addition, basic_subtraction, money_word_problem, measurement_word_problem, three_number_mixed, word_problem_context
	shorter_is_larger	basic_comparison, ordering, money_context, measurement_context, number_line, word_problem_context
exponents	whole_number_thinking	basic_addition, basic_comparison, subtraction, multiplication, word_problem, word_problem_context
	add_exponents_for_power_of_power	basic_power_of_power, with_coefficient, numerical_evaluation, product_of_powers, word_problem, word_problem_context
	distribute_exponent_over_addition	simple_two_term
	forget_exponent_on_coefficient	basic_power_of_power, larger_coefficients, negative_coefficient, multiple_variables, word_problem, word_problem_context
	multiply_base_by_exponent	simple_numeric, larger_exponent, word_problem_context
	multiply_exponents_when_multiplying_powers	basic_two_powers, three_powers, numerical_base, mixed_operations, word_problem, word_problem_context
factoring	negative_exponent_makes_negative	simple_numeric, larger_negative_exponent, fractional_base, word_problem_scientific, word_problem_finance, word_problem_context
	zero_exponent_equals_zero	basic_zero_exponent, expression_simplification, exponent_rules, polynomial_evaluation, word_problem, word_problem_context
	forget_gcf_first	default
	incomplete_factoring	gcf_then_difference_of_squares, gcf_then_trinomial, gcf_then_perfect_square, nested_factoring, word_problem
	negative_one_factor_forgotten	basic_difference_of_squares, coefficient_difference_of_squares, trinomial_factoring, gcf_then_pattern, comparison_problem, word_problem
sum_of_squares_factors	sign_errors_in_factoring	default
	sum_of_squares_factors	basic_sum_of_squares, coefficient_on_x, both_variables, larger_coefficients, word_problem

Table 17: Templates by malrule (Part 1 of 4).

Category	Malrule	Templates
fractions	add_numerators_denominators	default
	common_denominator_numerator	basic_addition, word_problem_context, three_fractions, word_problem_three_fractions, visual_representation, comparison_problem
	denominator_comparison_error	basic_two_fractions, three_fractions_ordering, word_problem_three_fractions, mixed_comparisons, word_problem_mixed, real_world_context, benchmark_comparison
	ignore_denominators	default
	keep_common_denominator_for_multiplication	basic_common_denominator, three_fractions, word_problem_three_fractions, mixed_numbers, word_problem_mixed, larger_denominators, word_problem
	multiply_across_for_division	basic_fraction_division, whole_number_divisor, word_problem_whole_divisor, whole_number_dividend, word_problem_whole_dividend, mixed_numbers, word_problem_mixed, word_problem
	multiply_rule_for_addition	default
natural_number_bias_numerator_only	basic_comparison, visual_models, real_world_context, multiple_choice, word_problem_multiple_choice, ordering, word_problem_ordering	
subtract_across	basic_subtraction, word_problem_pizza, three_fractions, word_problem_three_fractions, improper_fractions, word_problem_improper, word_problem_measurement	
functions	function_distributive_property	default
	function_notation_is_multiplication	default
	same_input_different_outputs_ok	ordered_pairs_set, table_format, graph_points, mapping_diagram, word_problem
	scalar_multiplication_inside_or_outside_same	quadratic_function, cubic_function, absolute_value, square_root, word_problem
geometry	count_net_perimeter_as_surface_area	cube_net, rectangular_prism_net, triangular_prism_net, pyramid_net, word_problem_context
	same_area_same_perimeter	default
	same_perimeter_same_area	default
	volume_formula_for_surface_area	rectangular_prism, cube, larger_dimensions, decimal_dimensions, word_problem
graphing	ignore_coordinate_signs	single_point_plotting, all_four_quadrants, distance_between_points, midpoint_calculation, word_problem
	reverse_coordinate_order	default
linear_equations	confuse_slope_and_intercept_roles	write_equation, identify_slope, identify_y_intercept, standard_form_identify, word_problem_context
	slope_direction_confusion	basic_positive_slope, basic_negative_slope, larger_magnitude, fractional_slope, comparison_problem, word_problem_context
	slope_is_delta_x_over_delta_y	basic_two_points, larger_coordinates, real_world_rate, negative_coordinates, mixed_quadrants
	y_intercept_always_positive	basic_slope_intercept, standard_form, point_slope_form, two_points_form, word_problem_context
measurement	inverted_conversion_factor	default
	wrong_conversion_factor	default

Table 18: Templates by malrule (Part 2 of 4).

Category	Malrule	Templates
multiplication_division	alignment_error_in_multi_digit	basic_2x2, word_problem, three_digit, money_context, array_model, three_digit_times_three, perimeter_then_area, multi_item_purchase, tiling_problem, calendar_calculation
	divide_larger_by_smaller_always	basic_smaller_dividend, fraction_result, decimal_friendly, word_problem_sharing, word_problem_measurement, money_sharing, probability_ratio, percentage_grade, ratio_comparison, scale_model
	division_makes_smaller	basic_decimal_division, fraction_division, measurement_context, sharing_context, rate_context, time_conversion, capacity_division, speed_calculation, recipe_scaling_up, unit_conversion
	forget_to_add_carried_number	basic_two_digit_times_one, three_digit_times_one, two_digit_times_two, money_context, word_problem, three_digit_times_two, area_calculation, total_cost_bulk, distance_calculation, array_larger
	multiplication_makes_bigger	basic_decimal_multiplication, fraction_multiplication, money_context, measurement_context, percent_discount, area_rectangle, probability_compound, compound_scaling, rate_distance, volume_box
negative_numbers	larger_absolute_value_always_wins	default
	multiplication_rule_for_addition	two_negatives_addition, positive_minus_negative, word_problem_temperature, word_problem_elevation, sequential_operations
	negative_swaps_operation	basic_arithmetic, word_problem_context, multi_step_expression, algebraic_context, comparison_problem
	negative_times_negative_negative	basic_multiplication, word_problem_temperature, word_problem_debt, equation_solving, pattern_completion, word_problem_multi_step, word_problem_mixed_operations, word_problem_distributive
two_negatives_always_positive	basic_subtraction, addition_negatives, word_problem_temp, word_problem_money, multi_step	
order_of_operations	addition_before_subtraction_always	simple_expression, money_transaction, temperature_change, elevation_change, score_tracking
	ignore_parentheses	basic_single_parentheses, nested_parentheses, multiple_parentheses, brackets_and_parentheses, word_problem
	multiplication_before_division_always	basic_expression, word_problem_sharing, word_problem_measurement, word_problem_money, multi_operation_chain
	pemdas_strictly_sequential	basic_mult_div, basic_add_sub, longer_expressions, mixed_operations, word_problem
percentages	strict_left_to_right	add_mult, sub_mult, add_div, sub_div, word_problem
	add_percentages_directly	default
	percent_equals_decimal	percent_to_decimal, decimal_to_percent, percent_in_calculation, fraction_to_percent, word_problem
percentage_as_index	basic_percentage, word_problem_context, percentage_increase, percentage_decrease, comparison_problem	
reverse_percentage_error	basic_reverse_relationship, percentage_increase, percentage_decrease, comparison_statements, word_problem	
radicals	add_under_common_root	basic_addition, basic_subtraction, three_radicals, word_problem_distance, mixed_operations, word_problem_addition, word_problem_subtraction, word_problem_three_radicals, word_problem_mixed
	distribute_square_root_over_addition	basic_sum_of_squares, pythagorean_context, algebraic_expression, subtraction_variant, non_perfect_square, word_problem_subtraction, word_problem_algebraic
	negative_outside_same_as_inside	basic_negative_outside, expression_form, in_equations, combined_operations, word_problem, word_problem_diverse, word_problem_combined_ops
	square_root_equals_plus_minus	basic_square_root, equation_vs_expression, negative_radicand_squared, expression_simplification, word_problem_context, word_problem_non_area, word_problem_equation_vs_expression
	square_root_is_divide_by_two	basic_perfect_square, larger_perfect_squares, area_of_square, equation_solving, pythagorean_word_problem, word_problem_non_geometric, word_problem_verification

Table 19: Templates by malrule (Part 3 of 4).

Category	Malrule	Templates
ratios_proportions	additive_instead_of_multiplicative	basic_scaling, recipe_scaling, speed_problems, similar_figures, mixture_problems, word_problem_context
	each_fraction_digit_is_ratio	basic_two_digit, three_digit_ratios, mismatched_lengths, with_zeros, word_problem_context
	ratio_as_division_only	part_to_part_ratio, recipe_scaling, paint_mixing, distance_rate, map_scale, word_problem_context
rounding	swap_ratios_or_units	default
	decimal_places_same_as_sig_figs	default
	leading_zeros_are_significant	basic_decimal_leading_zeros, measurement_context, scientific_notation_comparison, rounding_to_sig_figs, calculation_result, division_word_problem, unit_conversion_word_problem, density_calculation_word_problem, rate_calculation_word_problem, percentage_calculation_word_problem
	trailing_zeros_always_significant	basic_whole_number_trailing_zeros, measurement_with_units, scientific_notation_comparison, with_without_decimal, rounding_application, population_word_problem, distance_word_problem, rounding_result_word_problem, estimation_word_problem, large_scale_word_problem
scientific_notation	add_coefficients_when_multiplying	basic_multiplication, word_problem_context, multiple_step_calculation, compare_results, area_volume_calculation
	count_all_zeros_for_exponent	small_decimal_trailing_zero, large_number_trailing_zeros, word_problem_measurement, comparison_verification, mixed_zeros_decimal
	ignore_different_powers_of_ten	basic_addition, basic_subtraction, large_exponent_difference, negative_exponents, word_problem
statistics	wrong_exponent_sign	basic_conversion
	ignore_outliers_effect	basic_outlier, word_problem_context, multiple_outliers, symmetric_no_outlier, comparison_problem, word_problem_expanded
	mean_without_understanding	outlier_high, outlier_low, bimodal, skewed_right, best_measure_question, income_inequality_context, real_estate_market, environmental_data, daily_routine_outliers, word_problem_context
	mode_must_exist	basic_no_mode, larger_no_mode, small_no_mode, has_mode_control, comparison_problem, word_problem_context
subtraction	always_borrow_left	default
	borrow_from_bottom	basic_subtraction, word_problem_context, multi_step_problem, missing_number, comparison_problem
	borrow_no_decrement	default
	carry_ones_digit_instead_of_tens	basic_two_digit, larger_sums, three_numbers, money_context, word_problem
	decompose_by_place_value_label	basic_regrouped_tens, regrouped_ones, multiple_regroupings, four_digit, word_problem
	diff_0_n_equals_n	basic_single_zero, multiple_zeros, zero_in_ones_place, consecutive_zeros, word_problem
	no_column_limit	basic_two_digit, three_digit, multiple_carries, money_context, word_problem
	skip_equal	basic_skip_equal, multiple_equal_columns, word_problem_context, sequential_equal_digits, all_equal_digits
smaller_from_larger	basic_subtraction, money_context, measurement_context, three_digit, four_digit, missing_minuend, missing_subtrahend, comparison, multi_step, word_problem_result_unknown, word_problem_missing_minuend, word_problem_missing_subtrahend, word_problem_comparison	
word_problems	stops_borrow_at_zero	default
	subtract_smaller_from_larger_each_column	default
	include_all_numbers_given	shopping_cart, time_schedule, area_calculation, collection_combining, sharing_division, multi_item_purchase, multi_step_time, volume_calculation, perimeter_with_area, multi_step_collection, multi_step_division, mixed_operations

Category	Malrules	Templates	Avg	% Basic	% Word Prob
Multiplication & Division	5	50	10.0	8.0	8.0
Subtraction	11	47	4.3	23.4	21.3
Fractions	9	45	5.0	20.0	37.8
Radicals	5	37	7.4	16.2	43.2
Exponents	7	34	4.9	11.8	35.3
Decimals	5	30	6.0	23.3	26.7
Order of Operations	5	25	5.0	16.0	24.0
Negative Numbers	5	24	4.8	16.7	41.7
Statistics	3	22	7.3	9.1	18.2
Linear Equations	4	21	5.2	19.0	14.3
Rounding	3	21	7.0	14.3	47.6
Algebra	6	18	3.0	50.0	16.7
Factoring	5	18	3.6	22.2	16.7
Ratios & Proportions	4	18	4.5	16.7	16.7
Percentages	4	16	4.0	18.8	18.8
Scientific Notation	4	16	4.0	25.0	18.8
Absolute Value	3	12	4.0	25.0	16.7
Functions	4	12	3.0	16.7	16.7
Geometry	4	12	3.0	16.7	16.7
Word Problems	1	12	12.0	0.0	0.0
Graphing	2	6	3.0	16.7	16.7
Measurement	2	2	1.0	100.0	0.0
Total	101	498	4.9	18.3	24.5

Table 21: Template coverage across 22 mathematical categories. Each malrule has an average of 4.9 templates (ranging from 1 to 13), enabling diverse problem generation for each misconception. Template types include basic formulations (18.3%), real-world contexts (10.0%), word problems (24.5%), and structural variants (47.2%). The 498 templates span 334 unique template patterns, providing rich variety for cross-template generalization testing.

NCTM Strand	Categories	Malrules (%)
Number & Operations	4	54 (53.5%)
Algebra	3	37 (36.6%)
Geometry & Measurement	2	8 (7.9%)
Data & Modeling	1	4 (4.0%)
Total	10	101 (100%)

Table 22: Distribution of Malrules by NCTM Strand