





# MathFish : Evaluating Language Model Math Reasoning via Grounding in Educational Curricula

Li Lucy<sup>1,2</sup> Tal August<sup>1</sup> Rose E. Wang<sup>3</sup>  
Luca Soldaini<sup>1</sup> Courtney Allison<sup>4</sup> Kyle Lo<sup>1</sup>

<sup>1</sup>Allen Institute for AI <sup>2</sup>University of California, Berkeley  
<sup>3</sup>Stanford University <sup>4</sup>EdReports  
lucy3\_li@berkeley.edu kylel@allenai.org

## Abstract



To ensure that math curriculum is grade-appropriate and aligns with critical skills or concepts in accordance with educational standards, pedagogical experts can spend months carefully reviewing published math problems. Drawing inspiration from this process, our work presents a novel angle for evaluating language models’ (LMs) mathematical abilities, by investigating whether they can discern skills and concepts enabled by math content. We contribute two datasets: one consisting of 385 fine-grained descriptions of K-12 math skills and concepts, or *standards*, from Achieve the Core (ATC ) , and another of 9.9K math problems labeled with these standards (MathFish ). We develop two tasks for evaluating LMs’ abilities to assess math problems: (1) *verifying* whether a problem aligns with a given standard, and (2) *tagging* a problem with all aligned standards. Working with experienced teachers, we find that LMs struggle to tag and verify standards linked to problems, and instead predict labels that are close to ground truth, but differ in subtle ways. We also show that LMs often *generate* problems that do not fully align with standards described in prompts, suggesting the need for careful scrutiny on use cases involving LMs for generating curricular materials. Finally, we categorize problems in GSM8k using math standards, allowing us to better understand why some problems are more difficult to solve for models than others.

 **Code** [github.com/allenai/mathfish](https://github.com/allenai/mathfish)  
 **Dataset** [hf.co/datasets/allenai/mathfish](https://hf.co/datasets/allenai/mathfish)

## 1 Introduction

When assessing mathematical reasoning in large language models (LMs), a common approach is to test their problem solving abilities. Math is a popular domain for model evaluation (Cobbe et al., 2021; Hendrycks et al., 2021; Zhang et al., 2024), as problem instances can be designed to target

specific abilities. However, many datasets contain only coarse-grained information on what skills each problem assesses, such as general arithmetic or operation types (Hase et al., 2024). In practice, curricular experts’ categorizations of math are fine-grained, mapping materials to specific *skills*, such as multiplication procedures for fractions, or *concepts*, such as area and volume. Our work bridges this gap and examines a novel angle for evaluating LMs’ mathematical understanding. We ask, how well can models identify specific skills and concepts that students learn or practice when completing math problems?

First, we contribute English datasets of 9.9K human-written math problems (MathFish ) scaffolded by 385 K-12 U.S. curriculum standards in Achieve the Core (ATC ). These standards are informed by human learning progressions, and commonly used in real-world reviews of math content. In education, materials have focused *alignment* with a standard if they enable students to learn the *full intent* of concepts/skills described by that standard.<sup>1</sup> Identifying alignment can thus inform educators whether a set of materials adequately targets core learning goals for students.

Second, we provide a fine-grained assessment of LMs’ abilities to reason about math pedagogy, skills, and concepts, by asking them to recognize whether a standard aligns with a given problem (§4, §5). We experiment with two task formats: one in which we *verify* whether a single standard aligns with a problem, and another in which we *tag* each problem with any standards. Our experiments demonstrate that models achieve reasonable accuracy, but are not yet at expert-level performance across both task formats.

Third, to further illustrate the utility of these task formats, we apply the best-performing models and

<sup>1</sup><https://achievethecore.org/page/2730/aligned-instructional-practice>

prompting approaches on two case studies. In one, we work with K-12 math teachers with curriculum review experience to apply verification on LM-generated problems (§4.2). We find that a GPT-4 verifier tends to overestimate full standards alignment of generated problems compared to teachers. In the other, we tag GSM8k (Cobbe et al., 2021), a widely used grade school math dataset, with standards (§5.2). We find that GSM8k only covers around a third of all K-12 standards, and that LMs struggle more to solve problems tagged with higher grade levels’ standards.

As LMs are deployed in more real-world use cases, it is increasingly important to evaluate models with potential users. Throughout this paper, we work with curriculum specialists and teachers to center their voices when evaluating LM capabilities. These educational experts informed us that identifying math standards in curricular materials is time-intensive, and reviewing a set of published materials can take 6-8 months to complete. This motivates us to investigate whether models can support reviewers by combing through subtle differences across standards. Altogether, we hope that these datasets, along with the tooling we contribute for transforming data into task formats for LMs, can facilitate a more granular understanding of LM reasoning around skills and concepts in math content.

## 2 Related Work

**Math benchmarks for evaluating LMs.** Past work assessing mathematical reasoning capabilities of LMs have mainly focused on grade school arithmetic and algebra and problems that are easy to validate. These datasets may be written by annotators without formal pedagogical training (e.g. Cobbe et al., 2021; Patel et al., 2021; Amini et al., 2019; Ling et al., 2017; Miao et al., 2020), synthetically generated by LMs (Mitra et al., 2024; Liu et al., 2023a), and derived from advanced math competitions (Hendrycks et al., 2021). In contrast, our work characterizes real-world math curricula within an expansive, fine-grained taxonomy of math skills and concepts (§3). Concurrent work by Mishra et al. (2024) presents an approach for generating problems that align with math standards, but targets a limited set of easily validatable standards. Our work instead focuses on the task of labeling problems across 300+ comprehensive and challenging K-12 standards, and emphasizes ecological validity

by working with educators (§4.2). Earlier work in education has identified math standards in problems with models such as SentenceBERT (Li et al., 2024b), task-adapted BERT (Shen et al., 2021), and support vector machines (Karlovec et al., 2012). Our work re-envision this task into a format suitable for evaluating instruction-tuned LMs.

**LMs for improving math education.** LMs have increasing potential for improving math education through applications like automated feedback to educators (Jacobs et al., 2022; Demszky et al., 2023; Wang and Demszky, 2023), automated tutoring (Hobert and Meyer von Wolff, 2019; Wang et al., 2024), and lesson planning (Malik et al., 2024; Kasneci et al., 2023). Evaluating models with education domain experts before real-world deployment is crucial to best align with educational needs and mitigate potential harms (Patikorn et al., 2019; Li et al., 2024b; Lee et al., 2024). In our setting, we work with domain experts in a manner that centers real-world educational needs (§4.2).

**Evaluating LMs amid fine-grained differences.** Our emphasis on the granularity of models’ decisions relates to other work that focuses on small yet impactful differences in language. Examples include the generation of good distractors that sufficiently challenge question solvers (Gao et al., 2019; Zesch and Melamud, 2014), the mining of “hard negatives” to supervise models to discern positives and close negatives (Robinson et al., 2021; Kalantidis et al., 2020), and minimal modifications of model inputs that change ground truth labels (Gardner et al., 2020). In our case, our “hard negatives” and distractors are not artificially constructed, but instead originate from inter-standard relationships drawn by pedagogical experts.

## 3 Grounding Math Content in Expert-Designed Standards

### 3.1 Task Definition & Formats

What are task formats that both (1) evaluate LM ability to discern math skills or concepts in problems, as well as (2) reflect real-world usage patterns of LMs by pedagogical experts? In the initial phase of this work, we met with professional curriculum reviewers to better understand their processes and definitions for what does (or does not) constitute alignment to an educational standard. We identified two usage patterns in which these experts would employ language model assistance for performing

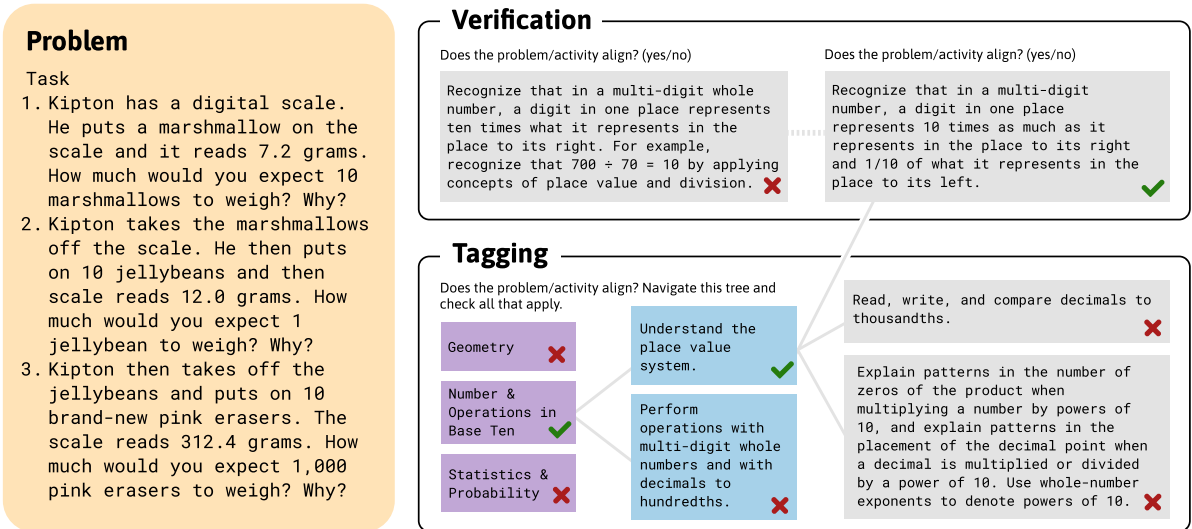


Figure 1: An example of a MathFish problem, along with domains ( $\mathcal{D}$ ), clusters ( $\mathcal{C}$ ), and standards ( $\mathcal{S}$ ) it does and does not align with. Solid lines indicate hierarchical relationships, while a dashed line links conceptually connected standards. In addition, this figure illustrates two task formats: verification (§4) and tagging (§5).

curricular alignment—*verifying* whether a problem aligns to a given standard and *tagging* a problem with all aligned standards.

When *verifying*, we check a single problem against a single standard. This may arise when one is confirming publishers’ claims of alignment (§4.1), or when evaluating whether LM-generated materials follow the standard indicated in prompt instructions (§4.2). This task format is structured as a binary yes/no question (Figure 1): does a problem fully align with a given standard? Taking inspiration from textual entailment (Dagan et al., 2005) and claim verification (Thorne et al., 2018), this format also allows us to investigate models’ sensitivity to narrowing differences in positive and negative standards when we perturb them in queries.

Though verification is useful for checking problem alignment with a *single* standard, reviewers are also often tasked with identifying *all* aligned standards for a given problem. In *tagging*, we take a single problem and select all aligned standards from a provided set of candidates. Prompting language models with hundreds of candidate standards risks hitting context limits or models getting “lost in the middle” (Liu et al., 2023b). Instead, we take advantage of a *domains-clusters-standards* tree structure that experts use to organize standards into a hierarchy; related standards are grouped as leaves on this tree (Figure 1). We define the tagging task as follows: given a math problem, (1) start at the top of the tree and select the best domain/s  $\mathcal{D}$  it teaches, (2) then traverse to each selected  $\mathcal{D}$ s’ subtrees, (3)

repeat selection for cluster/s  $\mathcal{C}$ , (4) stop traversal upon reaching standards  $\mathcal{S}$  at the leaves, and (5) tag the problem with selected  $\mathcal{S}$ . By traversing this tree, our tagging task format allows us to see whether models can make increasingly granular distinctions among adjacent concepts/skills.

Across these task formats, we experiment with three models: Mixtral-8x7B (Jiang et al., 2024), Llama-2-70b-chat (Touvron et al., 2023), and GPT-4-Turbo (Achiam et al., 2023). We unify them under a single wrapper, which handles input truncation and API calls (Appendix B).

### 3.2 Math Standards & Organization

To facilitate these experiments, we introduce two datasets: Achieve the Core (ATC ) describes math standards and their organization, and MathFish links standard labels to problems (Table 1). Both involve Common Core State Standards (CCSS), which offer fine-grained and comprehensive coverage of K-12 math skills/concepts (Porter et al., 2011).

ATC consists of CCSS standard labels (e.g. *4.NBT.A.1*) and descriptions (e.g. *Recognize that in a multi-digit whole number...*) from Achieve the Core’s coherence map, which captures how pedagogical experts characterize, organize, and categorize math.<sup>2</sup> ATC includes two types of relationships among standards: as leaves in a topical hierarchy, or as a graph of conceptual connections.

<sup>2</sup>[github.com/achievethecore/atc-coherence-map](https://github.com/achievethecore/atc-coherence-map)

Our experiments on verifying (§4) and tagging (§5) use these relationships to show how related standards that differ subtly can challenge models.

As described in §3.1, the CCSS hierarchy traverses the following levels from top to bottom: grade,  $\mathcal{D}$ ,  $\mathcal{C}$ , and  $\mathcal{S}$ .<sup>3</sup> In addition, ATC 🍏 includes *conceptual connections* between  $\mathcal{S}$  that cut across this hierarchy (Zimba, 2018). CCSS is designed to follow students’ learning progressions (Team, 2013), and recognizing these connections is central to enabling students to learn. That is, a problem intended to teach students addition would not jump directly to solving for variables. Figure 1, under *Verification*, shows an example of a conceptual connection: the 4th grade  $\mathcal{S}$  on the left (*Recognize that in a multi-digit whole number...*) progresses to the 5th grade  $\mathcal{S}$  on the right.

### 3.3 K-12 Math Problems

We evaluate models’ abilities to identify math skills and concepts using publisher-labeled data pulled from curricular websites. Publicly accessible open educational resources (OER) provide a rich test bed for evaluating models, as they are designed to cover nearly all  $\mathcal{S}$  within the grade levels they offer. We scrape pre-labeled problems from two OER that have been verified to be reputable by third party curriculum reviewers: Illustrative Mathematics and Fishtank Learning.<sup>4</sup> Each problem (e.g. the *Task* in Figure 1) in this combined dataset, which we refer to as MathFish 🐟, is a segment of these materials demarcated by  $\mathcal{S}$  labels, which we map onto ATC 🍏’s natural language descriptions. A problem in MathFish 🐟 can be labeled with multiple  $\mathcal{S}$ . Preprocessing details are in Appendix A.

We analyze models in text-only settings, replacing images with a dummy image token, and leave multimodal analyses for future work. We observed that many examples tend to include tables; we reformat HTML tables into Markdown or JSON based on models’ preferences during preliminary experiments (Appendix C). Though we contribute this entire scraped dataset of OER problems to facilitate future work, we evaluate models on a smaller evaluation set, which consists of 20% of all data. We do not evaluate models’ problem solving abilities on this data, and instead focus on identifying

<sup>3</sup>In high school, domains analogous to K-8 ones are called “categories”. More details on how we simplify complexities of CCSS like this one can be found in Appendix D.

<sup>4</sup>[illustrativemathematics.org/math-curriculum](http://illustrativemathematics.org/math-curriculum) and [fishtanklearning.org/about](http://fishtanklearning.org/about)

ATC: Math Standards Dataset 🍏			
Grade levels	K-12		
# of grades & domains, e.g. <i>4.NBT</i>	65		
# of clusters, e.g. <i>4.NBT.A</i>	147		
# of standards, e.g. <i>4.NBT.A.1</i>	385		
# of connections b/t $\mathcal{S}$	1,040		
Avg standard description length	36.23		
MathFish: Problems Dataset 🐟			
	IM	FL	Total
# of examples	19,570	2,206	21,776
# of labeled examples	7,735	2,188	9,923
Total words	4.5M	199K	4.7M
Avg problem length	230.51	90.20	216.30
Grade levels	K-12	3-11	K-12
# of unique standards	366	287	366

Table 1: An overview of standards from Achieve the Core (ATC 🍏) and curricular materials from MathFish 🐟, which combines Illustrative Mathematics (IM) and Fishtank Learning (FL). Lengths and word counts are based on white-spaced tokens.

math skills/concepts, as not all examples contain solutions, and some “problems” are designed to be group and/or hands-on activities.

### 3.4 Problems’ Alignment with Standards

Alignment is not always directly evident in problems’ language, and recognizing it requires a deeper understanding of cognitive processes. In Figure 1, each  $\mathcal{C}$  and  $\mathcal{S}$  relate to understanding the base ten number system, but differ in subtle ways. The purpose of the example problem on the left, based on the publisher’s description,<sup>5</sup> is to teach students how an understanding of the place value system can facilitate division and multiplication by ten. If the problem were instead intended to enable students to *Perform operations with multi-digit whole numbers and with decimals to hundredths*, a wider range of whole numbers than powers of ten would have been included in the problem for students to fully practice those procedural skills. As another example of misalignment, the problem in Figure 1 does not align with the left  $\mathcal{S}$  under *Verification*, as the problem involves place value understanding with decimals, which are not whole numbers. Thus, small differences in mathematical language have distinct consequences for student learning and assessment.

During evaluation, we assume that all  $\mathcal{S}$  not listed in MathFish 🐟 do not align with a prob-

<sup>5</sup>[tasks.illustrativemathematics.org/content-standards/tasks/1562](http://tasks.illustrativemathematics.org/content-standards/tasks/1562)

lem. We verify this assumption by having curriculum reviewers label a sample of problems paired with  $\mathcal{S}$  that are *not* listed to align with them in publishers’ materials, but are *similar* to positive labels. Here, we define “similar” as  $\mathcal{S}$  that are conceptually connected to positive labels in ATC 🍏’s map or are their same- $\mathcal{C}$  siblings. We recruited six teachers from a curriculum reviewing organization, which specializes in identifying CCSS alignment of materials. We paid these teachers approximately \$50 an hour, and within an allotted time frame of two weeks, they reviewed 136 pairs of problems with non-listed yet similar  $\mathcal{S}$ . Only 3 problem and  $\mathcal{S}$  pairs in this sample of assumed negatives were judged to actually be positives, thus estimating a ceiling for our assumption.

#### 4 Verifying Standards Alignment

This section focuses on our first task format, where models verify the alignment of individual standards against each problem. We first examine how models perform on MathFish 🐟 problems (§4.1), and then apply our best-performing verifier on generated problems (§4.2).

##### 4.1 Can LMs discern (mis)aligned standards?

**Prompt selection.** Since models can be sensitive to prompt design (Gonen et al., 2023; Sclar et al., 2023), we select prompts from a pool of 15 possible templates based on their performance in small, preliminary experiments (Appendix C.2). These prompts include standards’ descriptions to evaluate models’ language understanding abilities, and emphasize whether problems teach or enable students to learn a given concept or skill. Few-shot examples are sampled from a small exemplar pool spanning all grade levels (Appendix C.3). In the main text, we show results for each model’s top performing prompt template, but Appendix C.2 elaborates further on the ranges of performance scores we observed.

**Designing problem-standard pairs.** To set up the verification task, we take all gold problem-standard pairs in MathFish 🐟 to be positive. For negative labels, we sample them deliberately to vary their closeness to problems’ positive ones. For each problem, we assign five negative  $\mathcal{S}$  obtained via different sampling strategies. If a problem belongs to grade/s  $\mathcal{G}$  and domain/s  $\mathcal{D}$ , then we sample negative standards from  $\mathcal{G}$  or  $\mathcal{G}'$  and/or  $\mathcal{D}$  or  $\mathcal{D}'$ . Four sampling strategies emerge: ( $\mathcal{D}\mathcal{G}$ ,  $\mathcal{D}'\mathcal{G}$ ,

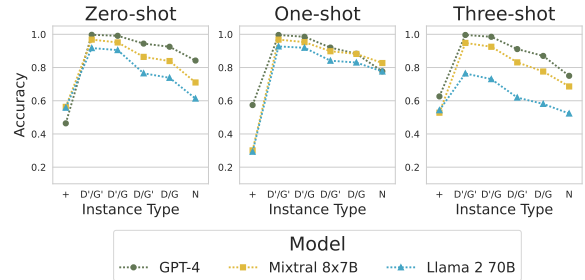


Figure 2: Verification accuracy when problems are paired with aligned standards (+) or with unaligned standards, ordered from left to right in increasing similarity to the positive standard ( $\mathcal{D}'\mathcal{G}' \rightarrow \mathcal{D}'\mathcal{G} \rightarrow \mathcal{D}\mathcal{G}' \rightarrow \mathcal{D}\mathcal{G} \rightarrow \mathcal{N}$ ). Language models have difficulty performing verification as standards become increasingly similar.

$\mathcal{D}\mathcal{G}'$ ,  $\mathcal{D}'\mathcal{G}'$ ). The fifth strategy involves sampling a negative standard that is conceptually connected to, or neighboring ( $\mathcal{N}$ ), positive ones in the ATC 🍏 coherence map. We expect that as sampled negatives moves closer to the positive standards in  $\mathcal{G}$  and  $\mathcal{D}$ , the difficulty of discerning true negatives will increase, thus lowering task performance.

**Experimental findings.** (i) As hypothesized, models’ accuracy decreases as negative examples become more conceptually similar to positive ones (Figure 2). Regardless, the best verifier is three-shot GPT-4, scoring the highest across all positive and negative pairs. (ii) Though Hase et al. (2024) showed that grade level relates to problem solving performance, our verification task does not have this relationship, as evidenced by insignificant Spearman’s  $\rho$  between grade level and F1, across all models and prompting approaches (Appendix C.5). This suggests that difficult aspects of our task extend beyond how easily a problem can be completed, posing a unique challenge for models. (iii) Finally, few-shot exemplars do not uniformly improve performance (Figure 2), nor does pairing task instances with exemplars in the same or nearby grades (Appendix C.3).

**Error analysis.** To obtain a closer look at models’ errors, we examine the language within particularly challenging  $\mathcal{S}$ . We find Mixtral and Llama-2’s false negatives are related to  $\mathcal{S}$  mentioning trigonometry (e.g. *sine*, *cosine*), while GPT-4 does not have this weakness (details in Appendix C.4). All three models also struggle with false positives mentioning proportions, ratios, and/or rates. These observations show how a standards-based evaluation framework can help identify model-specific

Error Type	Example $\mathcal{S}$ of a Generated Problem	Example Teacher Explanation
Problem goes beyond the learning stage of the $\mathcal{S}$ .	1.OA.A.2: <i>Solve word problems that call for addition of three whole numbers whose sum is less than or equal to 20...</i>	<i>Total number of beads (11+12) is greater than 20. [sic] and therefore beyond the boundary.</i>
Problem is too nonsensical or contains incorrect math.	G-CO.C.9: <i>Prove theorems about lines and angles. Theorems include: vertical angles are congruent; when a transversal crosses parallel lines, alternate interior angles are congruent and corresponding angles are congruent...</i>	<i>Directions are not correct and at best unclear. A transversal line cannot cross 2 parallel lines at 4 points. I think they mean angles, but where the angles are is unclear. And where is angle 3?</i>
Problem does not address some part of the $\mathcal{S}$ .	2.NBT.B.5: <i>Fluently add and subtract within 100 using strategies based on place value, properties of operations, and/or the relationship between addition and subtraction.</i>	<i>This is a good problem for practicing addition but it does not include subtraction.</i>
The solution is included as part of the problem setup.	7.G.B.4: <i>Know the formulas for the area and circumference of a circle and use them to solve problems; give an informal derivation of the relationship between the circumference and area of a circle.</i>	<i>The formulas for area and circumference are provided for students, therefore they are not able to demonstrate that they know the formulas to use independently [sic].</i>
Problem addresses other concepts/skills.	HSG-SRT.D.10: <i>Prove the Laws of Sines and Cosines and use them to solve problems.</i>	<i>This is not a Law of Sines/Cosines problem. This involved tangent and SOHCAHTOA.</i>

Table 2: The leftmost column shows common reasons for why generated problems have no or partial alignment, obtained via open coding of teachers’ explanations. Provided examples in each row are cases where GPT-4 judges a problem to be fully aligned, but teachers do not. Some  $\mathcal{S}$ ’s descriptions are shortened for brevity.

idiosyncrasies in a granular manner.

We also qualitatively observe misconceptions around the solution strategy targeted by a problem. For example, in one problem, students are asked to *Find the value of each expression mentally. 90-45, 270-45, 270-135, 360-135 ... How did this observation—that the first numbers are all multiples of 90—help you find the value of the differences?*<sup>6</sup> In other words, this problem is designed to leverage multiplication to find numerical differences. GPT-4’s response when verifying a positive  $\mathcal{S}$  related to multiplication instead claims that this problem *focuses on subtraction and mental math strategies... rather than on multiplication of whole numbers...* This example illustrates how identifying the concepts underlying problems requires recognizing how they are solved.

#### 4.2 Study 1: Verifying standard alignment of LM-generated problems

Diliberti et al. (2024)’s survey of teachers showed that some of the most common uses of AI in classrooms include generating materials such as assessments, lesson plans, and assignments. In addition, problem generation is a common task in AI & education research (e.g. Norberg et al., 2023; Zhou et al., 2023; Wang et al., 2021; Mishra et al., 2024; Shah et al., 2024), though generations are rarely evaluated by in-domain experts. In this section, we investigate the utility of GPT-4, our best verifier from §4.1, by applying it onto assessing *generated* math problems, working alongside K-12

<sup>6</sup>[im.kendallhunt.com/k5/teachers/grade-4/unit-7/lesson-2/lesson](https://im.kendallhunt.com/k5/teachers/grade-4/unit-7/lesson-2/lesson)

Model	Full (GPT-4)	Full (teachers)	Full + Partial (teachers)
Llama-2-70b	76%	19%	64%
Mixtral-8x7b	84%	35%	80%
GPT-4	96%	52%	85%

Table 3: Generated problems’  $\mathcal{S}$  alignment, as judged by a GPT-4-based verifier or by teachers.

math teachers to also obtain their verification judgments.

**Generating problems.** To obtain realistic generations, we asked six teachers from §3.3 to write example queries demonstrating how they would ask a model to generate math problems based on a  $\mathcal{S}$ . We used teachers’ suggested prompts to design 10 prompt templates in which we insert random  $\mathcal{S}$  labels and their descriptions to create 100 unique prompts (Appendix E). We input these prompts into Llama-2-70B, Mixtral-8x7B, and GPT-4 to generate a total of 300 problems.

**Collecting teachers’ annotations.** Sixteen teachers, recruited from the same curriculum reviewing organization as those in §3.3, then verified whether these generated problems align with the  $\mathcal{S}$  indicated in the original prompt. Teachers were distributed to grade levels based on their prior reviewing and teaching experience. Their annotations include the following labels: “not a math problem”, “no alignment”, “partial alignment”, or “full-intent alignment”, paired with written explanations. Since these problems are deliberately instructed to align with a  $\mathcal{S}$ , our generation process leads to many challenging, borderline cases. Each problem took 5-15

minutes for a teacher to judge, and we again paid teachers \$50 an hour. We compare teachers’ judgements to the best-performing GPT-4 verification approach (§4.1).

**Results.** From teachers’ judgements in Table 3, we see that most generated problems do not fully align with  $\mathcal{S}$  indicated in generation prompts. In addition, GPT-4 overestimates problem-standard alignment, and its estimated rates of full-intent alignment even exceed the combined percentage of full and partial ratings from teachers. Thus, this GPT-4 verifier skews more optimistic than teachers as a whole. Still, we observe agreement in the relative ranking of models for problem generation between teachers and our model verifier. In Table 2, teachers’ written explanations of why a generated problem does not enable a student to learn a given  $\mathcal{S}$  reveal the types of pedagogical considerations they make that are missed by problem generators and our GPT-4 verifier. Altogether, an LM verifier could be useful for estimating which models may generally generate better aligned problems, but may give less critical judgements than a team of curriculum reviewers.

## 5 Tagging Problems with Standards

As we introduced in §3, tagging navigates a hierarchical decision tree, where each branch presents models a problem and a list of  $\mathcal{D}$ ,  $\mathcal{C}$ , or  $\mathcal{S}$  descriptions. We shuffle the ordering of options in each level of the tree to avoid position bias (Zheng et al., 2023). The  $\mathcal{D}$  level consists of 12 options, while the  $\mathcal{C}$  level and  $\mathcal{S}$  level each have on average 13.7 options and 2.6 options, respectively. We also give models the option to also respond with “none” at each level, indicating that none of the listed options apply to an example problem. Prompt templates and our rationale behind their design can be found in Appendix D. In this section, we discuss experiments we run on MathFish problems (§5.1) and apply our best tagger on GSM8k (§5.2).

### 5.1 Can LMs traverse the tagging hierarchy?

**Prompt selection.** Like with verification, we again run small preliminary experiments using 15 possible prompt templates. In the main text, we present results for the top-performing prompt template for each model. We also experiment with one-shot and three-shot prompts, using the same problems we used to create few-shot exemplars for verification.

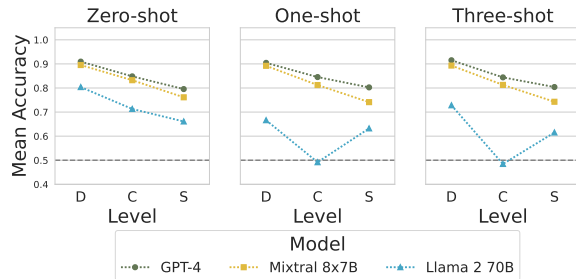


Figure 3: Average per-branch accuracy at each level ( $\mathcal{D}$ ,  $\mathcal{C}$ ,  $\mathcal{S}$ ) of the tagging tree during assisted traversal. The dashed line indicates a random baseline accuracy of 0.5. Stronger models decrease in performance when asked to make more granular decisions.

**Findings from assisted traversal setting.** We run several experiments that evaluate models at each level of the decision tree. Rather than have models traverse the tree on their own, we first evaluate in an “oracle”-assisted setting, where each level is presented with the assumption that the model has correctly chosen the correct branches in the previous level. We compute accuracy per branch based on the fraction of options correctly selected or not selected by the model. For example, if there are five options ( $A, B, C, D, E$ ), ground truth is  $B$ , and the model selects  $B$  &  $E$ , then accuracy on this branch is  $4/5$ .

Following our hypothesis, stronger LMs, including GPT-4 and Mixtral, tend to perform worse during assisted traversal as levels approach more fine-grained decisions (Figure 3). Yet for the weaker Llama-2, few-shot examples can hurt task performance rather than help, which may be due to how some models can struggle with *long* in-context examples more so than others (Li et al., 2024a). These model-specific findings also generalize to other prompt templates (Appendix D). Overall, the best performing model and prompt identified from these assisted traversal experiments is three-shot GPT-4 (mean accuracy = 0.850).

**Findings from self-guided traversal setting.** We then run three-shot GPT-4 on a more realistic tagging setup. Here, the model navigates the decision tree by relying on its own decisions in prior levels. Since models are allowed to respond “none” at any level of the tree, we are interested in seeing whether our best model and prompt pairing can recover during situations where it traverses down dead ends. We compute self-guided performance at the problem-level via two metrics: weak

Most Common $\mathcal{S}$	# (%)
4.OA.A.3. Solve multistep word problems posed with whole numbers and having whole-number answers using the <b>four operations</b> , including problems in which remainders must be interpreted...	250 (20.73%)
2.OA.A.1. Use <b>addition</b> and <b>subtraction</b> within 100 to solve one- and two-step word problems...	200 (16.58%)
3.OA.D.8. Solve two-step word problems using the <b>four operations</b> . Represent these problems using equations with a letter standing for the unknown quantity...	190 (15.75%)

Table 4: Excerpts from the top three most common  $\mathcal{S}$  tagged on GSM8k’s test set. References to  $+$   $-$   $\times$   $\div$  are **highlighted**, providing face validity for our tagger, three-shot GPT-4.

accuracy, where the model is “correct” if predicted  $\mathcal{S}$  overlaps with true  $\mathcal{S}$ , and exact accuracy, where the model is correct only if both sets are equal.

GPT-4 achieves an exact accuracy of only 0.048, and a weak accuracy of 0.502, based on the final predicted set of  $\mathcal{S}$  for each problem. In 59.3% of weakly accurate cases, predicted  $\mathcal{S}$  are a superset of gold ones, and GPT-4 tends to assign more  $\mathcal{S}$  per problem ( $M = 3.05$ ) than gold labels do ( $M = 1.66$ ). Exact accuracy varies across gold  $\mathcal{D}$ , ranging from 0.151 ( $\mathcal{D} =$  Counting & Cardinality) to 0.011 ( $\mathcal{D} =$  Functions). When GPT-4 traverses down a dead-end path in earlier levels, it is able to recover by predicting “none” at the  $\mathcal{S}$  level only 17.4% of the time. Though these metrics leave much room for improvement, predicted  $\mathcal{S}$  are often conceptually near ground truth. They have an average minimum distance of 1.9 edges from true  $\mathcal{S}$  in the ATC 🍏 coherence map, which is closer than an average minimum distance of 5.5 edges for random pairs of  $\mathcal{S}$ . 80.5% of predicted  $\mathcal{S}$  are in the same  $\mathcal{D}$  as gold  $\mathcal{S}$ , and 44.9% are in the same grade levels. Overall, our best LM setup can *approximate* where problems reside among  $\mathcal{S}$ , but pinpointing them is still out of reach.

## 5.2 Study 2: Tagging problems in GSM8k

Tagging problems with math concepts/skills can help document open math datasets and benchmarks. To illustrate this, we apply our task to GSM8k. GSM8k is a popular “grade school math” dataset used to evaluate models’ problem solving abilities (Cobbe et al., 2021). It has seen uptake by education-related papers (Jurenka et al., 2024), as well as in reports benchmarking LMs (e.g. Touvron et al., 2023; Chowdhery et al., 2023). We expect that most problems in GSM8k should align with  $\mathcal{S}$  pertaining to what Cobbe et al. (2021) describe as, “elementary calculations using basic arithmetic op-

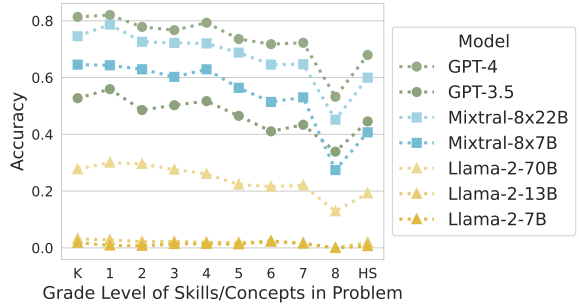


Figure 4: Models’ problem solving performance, based on the grade levels of  $\mathcal{S}$  tagged in problems.

erations ( $+$   $-$   $\times$   $\div$ )”. Thus, we aim to (1) confirm that GSM8k mostly covers arithmetic and (2) relate a disaggregation of problems by skills/concepts to models’ problem solving strengths and weaknesses.

**Experimental setup.** We apply our best performing tagging setup from §5 to GSM8k’s test set to approximate what  $\mathcal{S}$  it may cover. For math problem solving, we experiment with a range of model sizes within three families: GPT-4, GPT-3.5; Mixtral-8x7b, Mixtral-8x22b; Llama-2-7b, Llama-2-13b, and Llama-2-70b. We run these models on GSM8k’s test set using default settings in Eleuther AI’s evaluation harness (Gao et al., 2023).

**Results** Our tagger estimates that GSM8k covers 32.2% of all K-12  $\mathcal{S}$  and 56.1% of K-5  $\mathcal{S}$ , with the most common  $\mathcal{S}$  related to the four arithmetic operations (Table 4). So, though GSM8k may be linguistically diverse (Cobbe et al., 2021), it is not necessarily complete nor diverse in its coverage of grade school math skills and concepts. GSM8k is intended to be elementary-level (Cobbe et al., 2021), but some problems land in higher grade levels, e.g. linear equations in 8th grade. The three most frequent  $\mathcal{D}$  covered by GSM8k are Operations & Algebraic Thinking (36.0%), Expressions & Equations (25.8%), and Ratios & Proportional Relationships (21.8%).

Strikingly, across model families and sizes, problem solving performance on GSM8k relates to the grade levels of tagged  $\mathcal{S}$  (Figure 4). That is, though bigger models overall solve better than smaller ones, they suffer similar rates of performance degradation when they encounter problems involving higher grades’ skills and concepts. These trends follow a general intuition that problem solving difficulty should relate to “hardness” defined by grade level (Hase et al., 2024). Thus, even though



GSM8k problems do not originally come paired with grade level metadata, a noisy GPT-4 tagger is able to estimate grade-level difficulty using only natural language descriptions of math concepts and skills.

## 6 Conclusion

Our work investigates how well LMs can identify fine-grained skills and concepts that students learn when completing math problems. We contribute a dataset of 385 expert-designed mathematical standards layered with natural language descriptions and organizational metadata (ATC 🍏), a dataset of 9.9K standards-labeled problems drawn from reputable curricula (MathFish 🐟), and annotations of problem and standard pairs by experienced teachers. In addition, we provide tooling to transform data into tasks that assess LMs’ abilities to reason about mathematical language and concepts/skills. We find that though LMs still do not reach expert-level performance on verification, tagging, and generation tasks, they show promise and utility for assisting curriculum review and disaggregating math problem solving benchmarks.

## 7 Limitations

**Multimodality.** Math content is inherently multimodal (e.g. Lu et al., 2023), but our tasks focus on text-only data. The problems we gathered not only contain images, but sometimes embed interactive web applets. We plan to include images attached to problems in the final released form of our dataset to facilitate future work. The metadata attached to each example also includes the problem’s original url, in case others want to leverage other forms of information present in these online materials.

**Curriculum review.** Throughout the paper, we engage with teachers who professionally review curricular materials. During our process of working with them, we learned the ways in which their annotations for our paper may gloss over the complexities of how curriculum review occurs in practice. For example, when we generate problems based on standards, we include only one standard in the prompt. In reality, teachers may combine multiple standards in a single lesson, or use multiple problems to collectively target one standard. In addition, we asked teachers to individually annotate problem and standard pairs, and their annotations only consist of one pass over these materials (§3.3, §4.2). During actual reviews, teachers may collect

evidence of standards (mis)alignment individually, but later come together for careful deliberation, as flaws observed by one teacher may be missed by another. Finally, curriculum review typically evaluates materials for additional measures of quality beyond CCSS alignment, some of which we briefly discuss in Appendix E. We encourage future work to build on these investigations, especially as LMs become increasingly integrated into classrooms and educational technologies (Diliberti et al., 2024).

## 8 Ethical Considerations

LMs show promise as automated tools for gathering and/or suggesting standards (mis)alignment and assisting reviewers in their examination of materials. Though our paper aims to use LMs to automate the task of identifying standards alignment in curriculum, LMs’ role in curriculum review and creation processes should be a supporting, rather than leading, one. To design such tools, we believe that it is best to co-create with teachers and curriculum specialists.

In addition, the curriculum reviewers we worked with may not necessarily reflect the views of all teachers. Though they teach across the U.S., they are predominantly White and female (Appendix E). Our work also relies on Common Core, which is established in the U.S. and may not translate to pedagogical standards or practices in other socio-cultural contexts. A range of educators’ voices should be forefronted in research that intersects LMs and education.

ATC 🍏 is covered by a Creative Commons Public Domain Dedication License. Within MathFish 🐟, Illustrative Mathematics is licensed as CC BY 4.0, while Fishtank Learning component is licensed under Creative Commons BY-NC-SA 4.0. Both sources are intended to be OER, which is defined as teaching, learning, and research materials that provides users free and perpetual permission to “retain, reuse, revise, remix, and redistribute” for educational purposes.<sup>7</sup> The transformed versions of these materials as datasets is licensed under ODC-By 1.0, and our code to reproduce our experiments is licensed under Apache 2.0.

## 9 Acknowledgements

We thank the sixteen teachers from EdReports not only for their annotations, but also for their conversations which greatly informed our work. We

<sup>7</sup>[guides.library.columbia.edu/OER](https://guides.library.columbia.edu/OER)

also thank curriculum specialists at EdReports for patiently guiding us through the ecosystem of “real-world” math curriculum and curriculum review. Our work was generously funded by the Bill & Melinda Gates Foundation.

## References

- OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adela Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rameesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Valone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. [Gpt-4 technical report](#).
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [MathQA: Towards interpretable math word problem solving with operation-based formalisms](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily P Bonner. 2021. Practicing culturally responsive mathematics teaching. *Mathematics Teacher: Learning and Teaching PK-12*, 114(1):6–15.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.

- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. [The PASCAL recognising textual entailment challenge](#). In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg, Springer-Verlag.
- Dorottya Demszky, Jing Liu, Heather Hill, Dan Jurafsky, and Chris Piech. 2023. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.
- Melissa Kay Diliberti, Heather L Schwartz, Sy Doan, Anna Shapiro, Lydia R Rainey, and Robin J Lake. 2024. Using artificial intelligence tools in K–12 classrooms.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. [A framework for few-shot language model evaluation](#).
- Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. [Generating distractors for reading comprehension questions from real examinations](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models' local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Hila Gonen, Srinu Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. [Demystifying prompts in language models via perplexity estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.
- Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. 2024. [The unreasonable effectiveness of easy training data for hard tasks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7002–7024, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Sebastian Hobert and Raphael Meyer von Wolff. 2019. Say hello to your new automated tutor—a structured literature review on pedagogical conversational agents. In *Proceedings of the 14th International Conference on Wirtschaftsinformatik*. Siegen.
- Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. [Mixtral of experts](#). *ArXiv*, abs/2401.04088.
- Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, Roni Rabin, Jasmin Rubinovitz, Amit Pitaru, Mac McAllister, Julia Wilkowsky, David Choi, Roe Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbour, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Gorur, Svetlana Grant, Rachel Hashimshoni, Laura Weidinger, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Chris Summerfield, James An, Pierre-Alexandre Kamienny, Abhinit Mohdi, Theofilos Strinopoulos, Annie Hale, Wayne Anderson, Luis C. Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. 2024. [Towards responsible development of generative AI for education: An evaluation-driven approach](#). *Preprint*, arXiv:2407.12687.

- Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809.
- Mario Karlovec, Mariheida Córdova-Sánchez, and Zachary A Pardos. 2012. Knowledge component suggestion for untagged content in an intelligent tutoring system. In *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11*, pages 195–200. Springer.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Gloria Ladson-Billings. 2021. [Three decades of culturally relevant, responsive, & sustaining pedagogy: What lies ahead?](#) *The Educational Forum*, 85(4):351–354.
- Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Math multiple choice question generation via human-large language model collaboration. *17th International Conference on Educational Data Mining*.
- Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. 2024a. [Long-context llms struggle with long in-context learning](#). *Preprint*, arXiv:2404.02060.
- Zhi Li, Zachary A Pardos, and Cheng Ren. 2024b. Aligning open educational resources to new taxonomies: How AI technologies can help and in which scenarios. *Computers & Education*, 216:105027.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.
- Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023a. TinyGSM: achieving > 80% on GSM8k with small language models. *arXiv preprint arXiv:2312.09241*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.
- Rizwaan Malik, Dorna Abdi, Rose E. Wang, and Dorottya Demuszky. 2024. [Scaling high-leverage curriculum scaffolding in middle-school mathematics](#). In *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S '24*, New York, NY, USA. Association for Computing Machinery.
- Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. [A diverse corpus for evaluating and developing English math word problem solvers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.
- Shubhra Mishra, Gabriel Poesia, Belinda Mo, and Noah D. Goodman. 2024. [MathCAMPS: Fine-grained synthesis of mathematical problems from human curricula](#). *Preprint*, arXiv:2407.00900.
- Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-Math: Unlocking the potential of SLMs in grade school math. *arXiv preprint arXiv:2402.14830*.
- Kole Norberg, Husni Almoubayyed, Stephen E Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steven Ritter. 2023. Rewriting math word problems with large language models. In *AIED23: Artificial Intelligence in Education, Empowering Education with LLMs Workshop*.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Thanaporn Patikorn, David Deisadze, Leo Grande, Ziyang Yu, and Neil Heffernan. 2019. Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 396–405. Springer.
- Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. 2011. [Common Core Standards: The new U.S. intended curriculum](#). *Educational Researcher*, 40(3):103–116.
- Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. [Contrastive learning with hard negative samples](#). In *International Conference on Learning Representations*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

- Vedant Shah, Dingli Yu, Kaifeng Lyu, Simon Park, Nan Rosemary Ke, Michael Mozer, Yoshua Bengio, Sanjeev Arora, and Anirudh Goyal. 2024. [AI-assisted generation of difficult math questions](#). Preprint, arXiv:2407.21009.
- Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. 2021. Classifying math knowledge components via task-adaptive pre-trained BERT. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22*, pages 408–419. Springer.
- Common Core Standards Writing Team. 2013. Progressions for the Common Core State Standards in mathematics. *Institute for Mathematics and Education, University of Arizona*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Rose Wang and Dorothea Demszky. 2023. [Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.
- Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorothea Demszky. 2024. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Zichao Wang, Andrew Lan, and Richard Baraniuk. 2021. [Math word problem generation with mathematical consistency and problem context constraints](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5986–5999, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Torsten Zesch and Oren Melamud. 2014. [Automatic generation of challenging distractors using context-sensitive inference rules](#). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Baltimore, Maryland. Association for Computational Linguistics.
- Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Zihao Zhou, Maizhen Ning, Qiufeng Wang, Jie Yao, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2023. [Learning by analogy: Diverse questions generation in math word problem](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11091–11104, Toronto, Canada. Association for Computational Linguistics.
- Jason Zimba. 2018. [A graph of the content standards](#). Technical report.

## A Data Collection and Preprocessing

### A.1 Scraping MathFish

**Illustrative Math.** Illustrative Math (IM) segments materials across their website based on their intended use in classrooms. We pull problems from the following parts of their website: tasks, lessons, centers, practice, and modeling prompts.  $\mathcal{S}$  are listed with varying relations to content, e.g. a problem is “Building Towards” a  $\mathcal{S}$ , if it is foundational but not at the level of a standard. We use the relations “Addressing” and “Alignment” as ground truth positive labels for  $\mathcal{S}$  alignment. The dataset we contribute includes all problems and label types, in case future work wishes to examine discerning how problems that are “Building On” a  $\mathcal{S}$  may differ from those that are “Addressing” or “Building Towards” it.

**Fishtank Learning.** We obtain problems from Fishtank Learning (FT) from lessons listed within each unit. FT includes two types of labels: “Core Standards” and “Foundational Standards”, the latter of which is similar to IM’s “Building Towards.” We use  $\mathcal{S}$  listed as “Core Standards” as ground truth positive labels for alignment. We again include all problems and label types in the dataset that accompanies our paper.

Table format	Mixtral 8x7B	Llama-2 70B	GPT-4-turbo
reStructuredText	0.855	0.666	0.882
markdown	0.871	<b>0.710</b>	<b>0.880</b>
json	<b>0.885</b>	0.696	0.869
html	0.853	0.664	0.861

Table 5: F1 scores during preliminary verification experiments to determine models’ table formatting preferences.

## A.2 Label Preprocessing

**Standardization.**  $\mathcal{S}$  can be written in a variety of ways by educators and curricular materials (e.g. *HSS-MD.B.5* is the same as *S-MD.5*). We standardize these  $\mathcal{S}$  labels so that we can link them across datasets, and use the label version present in Achieve the Core (ATC 🍏) as the canonical label.

**Inheritance.** Not all labels present in OER materials are at the  $\mathcal{S}$  level. If a  $\mathcal{C}$  is listed for a problem, then we infer that the problem aligns with all  $\mathcal{S}$  within that  $\mathcal{C}$ . Similarly, if a sub-standard (e.g. *F-IF.C.7a*, *F-IF.C.7b*) is listed for a problem, we assume it aligns with its parent  $\mathcal{S}$ .

## B Model Wrapper

For both verification and tagging, we unify all models under a single model wrapper to keep prompting consistent across them. We use the TogetherAI API<sup>8</sup> and OpenAI API<sup>9</sup> for model access. In cases where a prompt exceeds a model’s context window, we truncate the problem description in the prompt, but retain the entirety of  $\mathcal{S}$  descriptions. During experiments, all models are given 3 retries for incorrect response formatting (e.g., not including a *yes* or *no* in the verification task format). Retries call the model again with no additional context. Models were run using their default temperature and maximum context window. In total, we spent less than \$5k on API calls.

## C Verification

### C.1 Table Formatting

Web-scraped math problems sometimes include tables. We first experimented with different table formatting styles in one fixed prompt template: HTML, JSON, Markdown, and reStructuredText. We evaluate on a random sample of 500 verification instances, which consist of OER problems from

our evaluation set paired with positive labels or negative labels sampled from  $\mathcal{D}'\mathcal{G}'$ . Using a fixed prompt template, we find that Llama-2 and GPT-4 prefer Markdown tables, while Mixtral prefers JSON (Table 5). We retain these table preferences for all further experiments in our paper, including those for tagging.

### C.2 Prompts

Three authors collaboratively wrote a pool of 15 prompt templates  $\{\mathcal{P}_i \mid i = 1, 2, 3, \dots, 15\}$  which vary in phrasing. These templates are designed to emphasize whether problems teach or enable students to learn a given concept or skill. Their paraphrases were informed by language occurring in resources that discuss Common Core alignment, especially “full intent” or “focused” alignment (Figures 10-15).<sup>10</sup> We again evaluate on a random sample of 500 examples as we did in §C.1.

Models vary in performance across  $\mathcal{P}_i$  during these small preliminary experiments, though there is some overlap across models’ top three. Across all 15 prompt templates, we observe F1 ranges of 0.605-0.838 for Llama 2, 0.778-0.895 for Mixtral, and 0.778-0.913 for GPT-4. Each models’ top-3 prompts are:  $\mathcal{P}_4$  (Figure 11),  $\mathcal{P}_{10}$  (Figure 14), and  $\mathcal{P}_{15}$  (Figure 15) for Llama 2,  $\mathcal{P}_1$  (Figure 10),  $\mathcal{P}_5$  (Figure 12),  $\mathcal{P}_{15}$  (Figure 15) for Mixtral, and  $\mathcal{P}_1$  (Figure 10),  $\mathcal{P}_4$  (Figure 11),  $\mathcal{P}_7$  (Figure 13) for GPT-4.

Then, we ran these top-3 performing prompt templates on the full evaluation set, across all five types of negative labels as described in the main text §4. Figure 5 shows the variation we obtained in performance across prompts, though general trends reflect the same conclusions as Figure 2 in the main text. That is, as negative labels are more similar to positive ones, verification accuracy decreases, and different model and prompt pairings trade off false positives and false negatives differently.

### C.3 Few-Shot Exemplars

We experiment with zero-shot, one-shot, and three-shot prompts. From problems outside of our evaluation set, we pulled one problem from grades K-8 and three from HS. We pair these problems with one positive label and one negative one, which is a randomly selected, conceptually similar neighbor to the problem’s positive labels in

<sup>8</sup><https://docs.together.ai/docs/quickstart>

<sup>9</sup><https://platform.openai.com/docs/overview>

<sup>10</sup>[https://curriculum.illustrativemathematics.org/MS/teachers/design\\_principles.html](https://curriculum.illustrativemathematics.org/MS/teachers/design_principles.html), <https://achievethecore.org/page/1118/coherence-map>

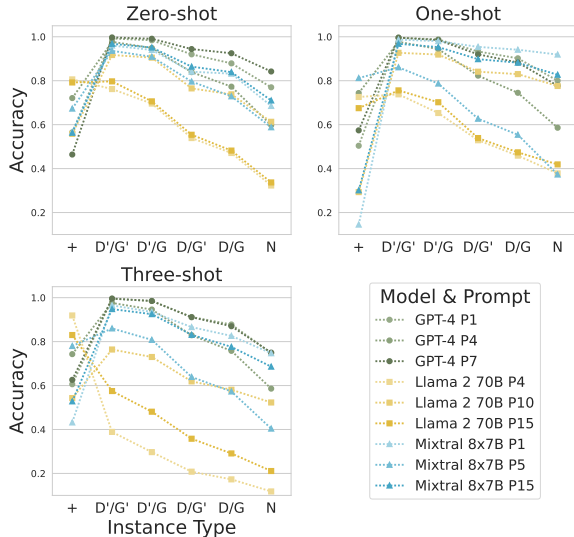


Figure 5: Verification accuracy when problems are paired with aligned standards (+) or with unaligned standards, ordered from left to right in increasing similarity to the positive standard ( $D'G' \rightarrow D'G \rightarrow DG' \rightarrow DG \rightarrow N$ ). Language models have difficulty performing verification as standards become increasingly similar.

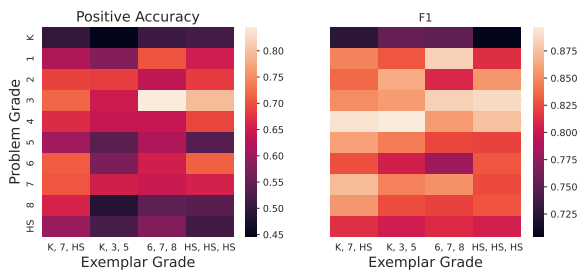


Figure 6: Three-shot verification performance, disaggregated across problems’ grade levels and few-shot exemplars’ grade levels. We did not find a clear relationship between problem grade and exemplar grade.

the ATC 🍏 map. We write Chain-of-Thought-like explanations for these problem and  $\mathcal{S}$  pairs to create few-shot exemplars to insert into prompts (e.g. *Example {i}* in Figures 10-15 would be repeated for each exemplar). An example of a few-shot exemplar:

### Problem:

You’re mixing ingredients for cookies. The recipe says to combine 6 tablespoons, or  $\frac{3}{4}$  stick, of butter with 1 cup of sugar. You accidentally mix in a whole stick of butter (8 tablespoons) with the cup of sugar. How can you fix this?

### Standard description:

Use ratio and rate reasoning to solve real-world

and mathematical problems, e.g., by reasoning about tables of equivalent ratios, tape diagrams, double number line diagrams, or equations.

### Answer:

yes

### Explanation:

This problem solves a real-world problem involving mixing ingredients for cookies. A student doing this problem would need to reason about equivalent ratios of butter amounts to sugar amounts.

For one-shot prompts, we randomly select an exemplar problem and  $\mathcal{S}$  pair from any grade. For three-shot prompts, we select exemplars that span wide and narrow grade ranges:  $\{(K, 7, HS), (K, 3, 5), (6, 7, 8), (HS, HS, HS)\}$ . We initially hypothesized that models may perform better on problems accompanied by few-shot exemplars in similar grade spans, but we did not confirm this hypothesis with our exemplar pool (Figure 6).

## C.4 Error Analysis

In the main paper, we discuss false negative and false positive patterns among models by referring to the language within particularly challenging  $\mathcal{S}$ . To do this analysis, we tokenize  $\mathcal{S}$  using Bling Fire, and remove tokens that are less than 2 characters long, are nltk English stopwords, or appear in less than 5  $\mathcal{S}$ . Table 6 shows the results of this analysis.

## C.5 Performance Across Grade Levels

Intuitively, higher grade levels may suggest lower verification performance, based on known measures of problem “hardness” (Hase et al., 2024). However, Figure 7 shows a lack of a consistently positive trend between F1 (calculated over all positive and negative examples) and problems’ grade level. For all 27 combinations of model and prompting approaches, we did not observe a significantly positive Spearman  $\rho$  between performance and grade level, where significance is measured as  $p < 0.05$  with Bonferroni correction.

## D Tagging

### D.1 Prompts

We write 15 possible prompt templates  $\{Q_i \mid i = 1, 2, 3, \dots, 15\}$  and run models on 500 random instances of a simple toy task to filter out catastrophic templates. In this toy task, we provide models an

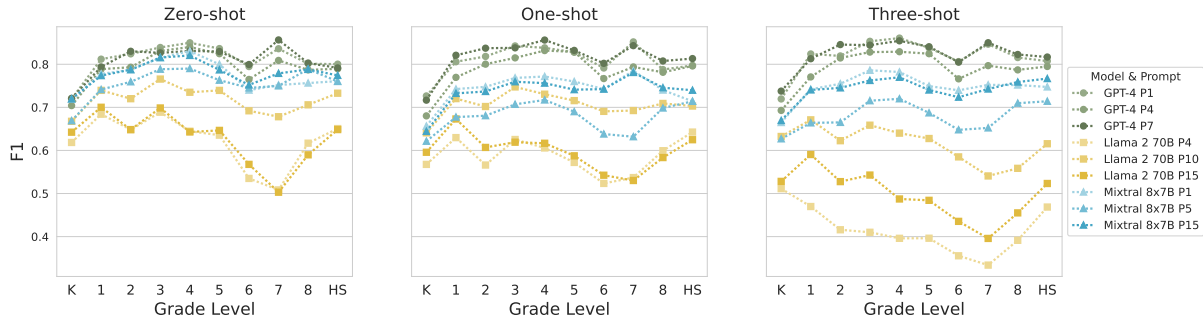


Figure 7: Verification performance across all positive and negative examples, separated by grade level on the  $x$ -axis. Each model is shown with its top-3 prompt templates, which were identified during preliminary verification experiments.).

Error	Words in $\mathcal{S}$ (Error Rate)
Model: three-shot GPT-4	
FP	proport (0.26), ratio (0.24), rate (0.24), input (0.22), verbal (0.2), quantiti (0.2), descript (0.19), relationship (0.19), assign (0.18), equival (0.18)
FN	invers (0.74), argument (0.69), half (0.68), unlik (0.61), partit (0.61), similar (0.6), congruenc (0.6), prove (0.6), cylind (0.58), first (0.56)
Model: one-shot Mixtral	
FP	proport (0.27), extend (0.26), previou (0.25), featur (0.25), reason (0.22), mathemat (0.22), assess (0.21), quantiti (0.21), diagram (0.2), neg (0.2)
FN	sine (1.0), cosin (1.0), polynomi (0.97), invers (0.95), half (0.91), congruent (0.91), similar (0.9), definit (0.89), transform (0.88), argument (0.87)
Model: one-shot Llama-2	
FP	proport (0.35), person (0.28), rate (0.28), extend (0.27), previou (0.27), ratio (0.26), per (0.25), quantiti (0.25), hour (0.24), descript (0.24)
FN	ident (1.0), sine (0.91), cosin (0.91), name (0.89), count (0.88), true (0.88), 1,000,000 (0.88), need (0.88), partit (0.87), origin (0.85)

Table 6: Stemmed words in  $\mathcal{S}$  that are most difficult for each model’s best prompting approach, with error rates in parentheses. FN = false negative, FP = false positive.

OER problem and 5 random  $\mathcal{S}$  descriptions, and models are asked to select positive labels hidden among these options.

Several elements in tagging prompt templates (indicated in curly brackets in Figures 16-23) vary depending on the level of the tagging decision tree. We outline these elements here:

#### Domain

- `relation_definition = ‘`
- `level = ‘topics’`
- `Level = ‘Topic’`
- `relation = ‘teaches’`
- `options =` These map onto K-8 domains and HS categories, but are not a one-to-one mapping. Some high school (HS) categories are equivalent or similar to a domain in K-8, and some differences in K-8 domains are difficult to explain a brief description at the domain-level. Thus, a “domain” in our paper sometimes groups multiple actual CCSS

domains/categories. We mostly retain the original CCSS K-8 domains and HS categories, but make exceptions for the following: we group OA (Operations & Algebraic Thinking), EE (Expressions & Equations), and A (HS Algebra) into *Operations & Algebra*, S (HS Statistics & Probability) and SP (K-8 Statistics & Probability) to *Statistics & Probability*, and finally NS (K-8 The Number System) and N (HS Number and Quantity) to *Number Systems and Quantity*. Since CCSS and ATC 🍏 do not provide brief descriptions of domains, we worked with a curriculum specialist to write these descriptions of each  $\mathcal{D}$  option (Figure 9).

#### Cluster

- `relation_definition = ‘`
- `level = ‘mathematical concepts/skills’`
- `Level = ‘Mathematical concepts/skill’`
- `relation = ‘teaches’`
- `options =` Options are natural language descriptions of clusters, obtained from ATC 🍏.

#### Standard

- `relation_definition = ‘A problem or activity aligns with a standard if it can enable students to learn the full intent of the concepts and skills outlined in the standard’s description.’`
- `level = ‘standards’`
- `Level = ‘Standard’`
- `relation = ‘aligns with’`
- `options =` Options are natural language descriptions of standards, obtained from ATC 🍏.

We observed models struggling to follow response formatting instructions in early experiments.



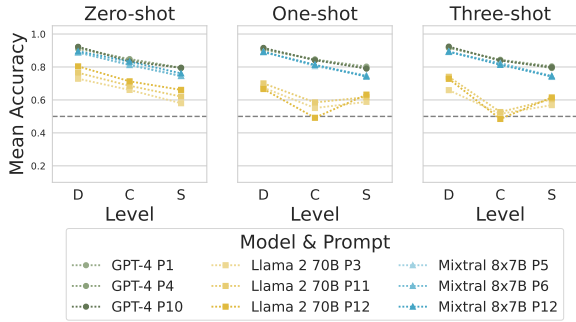


Figure 8: Average per-branch accuracy at each level ( $\mathcal{D}$ ,  $\mathcal{C}$ ,  $\mathcal{S}$ ) of the tagging tree during assisted traversal. The dashed line indicates a random baseline accuracy of 0.5. Stronger models decrease in performance when asked to make more granular decisions.

Thus, our prompt paraphrases also vary specifications around response format, e.g. comma-separated selected options (Figure 16). In addition,  $\mathcal{P}_7$ - $\mathcal{P}_{15}$  suggested walking through the steps of solving or completing a given problem (e.g. Figure 22). To rank  $\mathcal{Q}_i$  per model, we first calculate the rate to which models format responses correctly, and break ties based on models’ weak accuracy, where responses are correct if predictions overlap with gold labels.

Models, like with verification, vary in performance across all 15  $\mathcal{Q}_i$ . On our toy task, we observe exact accuracy ranges of 0.046-0.332 for Llama 2, 0.304-0.542 for Mixtral, and 0.586-0.75 for GPT-4. This performance ordering of models parallels the one obtained for verification, though the gaps among models and prompts are greater here. Some  $\mathcal{Q}_i$  favored by one model are highly detrimental to another, and asking a model to walk through doing a problem does not necessarily improve performance. Each models’ top-3 prompts are:  $\mathcal{Q}_3$  (Figure 17),  $\mathcal{Q}_{11}$  (Figure 22),  $\mathcal{Q}_{12}$  (Figure 23) for Llama 2,  $\mathcal{Q}_5$  (Figure 19),  $\mathcal{Q}_6$  (Figure 20),  $\mathcal{Q}_{12}$  (Figure 23) for Mixtral, and  $\mathcal{Q}_1$  (Figure 16),  $\mathcal{Q}_4$  (Figure 18),  $\mathcal{Q}_{10}$  (Figure 21) for GPT-4.

We ran these top-3 performing prompt templates for each model on the full evaluation set, and find that the trends we describe in the main text’s Figure 3 generalize to other top-performing prompts (Figure 8). That is, as stronger models are asked to make more granular decisions from  $\mathcal{D}$  to  $\mathcal{C}$  to  $\mathcal{S}$ , their performance decreases.

## D.2 Few-Shot Exemplars

We again wrote explanations for  $\mathcal{D}$ -,  $\mathcal{C}$ -, and  $\mathcal{S}$ -level tagging decisions to create exemplars, and sampled one-shot and three-shot exemplars in the same manner as we did with verification. An example of an exemplar at the  $\mathcal{S}$ -level:

### Problem:

Task

Cruz and Erica were both getting ready for soccer. Cruz ran 1 lap around the school.

Erica ran 3 laps around the playground.

Erica said,

I ran more laps, so I ran farther.

Cruz said,

4 laps around the school is 1 mile, but it takes 12 laps around the playground to go 1 mile. My laps are much longer, so I ran farther.

Who is right? Draw a picture to help you explain your answer.

### Options:

A. Explain why a fraction  $a/b$  is equivalent to a fraction  $(n \times a)/(n \times b)$  by using visual fraction models, with attention to how the number and size of the parts differ even though the two fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions. Grade 4 expectations in this domain are limited to fractions with denominators 2, 3, 4, 5, 6, 8, 10, 12, and 100.

B. Compare two fractions with different numerators and different denominators, e.g., by creating common denominators or numerators, or by comparing to a benchmark fraction such as  $1/2$ . Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols  $>$ ,  $=$ , or  $<$ , and justify the conclusions, e.g., by using a visual fraction model. Grade 4 expectations in this domain are limited to fractions with denominators 2, 3, 4, 5, 6, 8, 10, 12, and 100.

### Answer:

A, B

### Explanation:

This task asks students to compare two fractions that are equal. The two fractions have different numerators and denominators (e.g.  $1/4$  and  $3/12$ ). The students need to explain why the two fractions

		Count	Percentage
<b>Race</b>	Black/African American	2	12.50%
	White (Caucasian)	14	87.50%
<b>Gender</b>	Female	13	81.25%
	Male	1	6.25%
	Prefer not to answer	2	12.50%
<b>Region</b>	West	2	12.50%
	Southeast	3	18.75%
	Northeast	4	25.00%
	Midwest	7	43.75%
<b>Yrs in Education</b>	0-10 years	4	25.00%
	11-20 years	3	18.75%
	21-30 years	7	43.75%
	30+ years	2	12.50%

Table 7: Teachers’ experiential and sociodemographic backgrounds.

are equal by drawing a picture, which involves using visual fraction models.

## E Generated Problems Study

We worked with sixteen U.S.-based, K-12 teachers from a curriculum reviewing organization to annotate pairs of problems and standards (§3.3, §4.2). These teachers differ from teachers one may recruit from other sources (e.g. Upwork) in that they have prior experience reviewing Common Core alignment of published curricular materials. We worked with three cohorts of 4-6 of teachers, through face-to-face virtual meetings to provide an initial set of instructions, and then communicated with them throughout the annotation process to clarify questions. These teachers were explicitly told that their annotations and explanations, but no identifying information, would be released as part of this project. Each teacher was paid a stipend that averaged \$50 an hour. Most of these teachers are White women, and they span regions across the U.S. and have varying lengths of experience in education (Table 7).

### E.1 Creating Prompt Templates

First, we asked a subset of teachers (our first cohort) to write prompts that reflect the following: *Pretend you have 1-3 standards (not the MP standards) in mind that you want to teach. How would you ask a model to generate a math problem based on these standard/s?* Here, “MP standards” refer to mathematical practice standards, while our work focuses on CCSS math *content* standards.

We found that teachers usually refer to standards by their labels, e.g. *4.NBT.A.1*, when searching for curricular resources, rather than specify full

Prompt Templates for Generating Problems
Create a problem for {standard}. This problem will be for a special education resource student.
Create a math problem for {standard}. This problem will need to be simplified to support diverse learners in the resource room.
Create a math problem that uses {standard}.
please generate a possible quiz item aligned to common core SS for Math for {standard}
Please generate an item aligned to CCSS math {standard} for students to use to practice with.
my math class is very diverse, with many nationalities, races, languages, and genders represented along with diverse family structures. Please generate a culturally competent practice item for use in an assessment with the standard {standard}.
Please generate a practice problem aligned to {standard}.
Create a multi-step word problem aligned to {standard} with at least 3 steps to solve. Make the the problem engaging and relevant for kids who are interested in culturally responsive, real-life scenarios that are fun. Make it easily adjustable for me to change some words/numbers around or edit as needed.
Can you create a problem aligned to {standard} that requires students conceptually understand this standard? Make one that is engaging, open, and culturally responsive.
create a problem aligned to {standard} in kid friendly language or broken down for ELL students to understand.

Table 8: Prompt templates inspired by teachers’ suggestions, which we used for generating math problems in §4.2.

standard descriptions. In addition, they do not always specify in their prompts that the standards are CCSS standards, rather than some other set of standards that may use similar labels (e.g. state-specific standards). Thus, at the end of each prompt template, we appended each CCSS standard’s natural language description to better guide models, e.g. Standard {label}: {description}.

Altogether, teachers wrote around 20 prompts, though due to variation in teachers’ prior experience with LMs, some prompts were not suitable as instruction-like inputs, e.g. *I’d like to see if AI could quickly generate sets of problems with specific root types, such as integers, fractions, etc both with the coefficient of  $x^2$  as 1 and with it as an integer*. Thus, we only select only a subset of all proposed prompts, especially ones that could be minimally edited into prompt templates (Table 8). We remove extraneous, standard-specific information, e.g. *My students need to prove they can add and subtract within 10 using models*, to make prompts generalizable for nearly all standards. Teachers’ suggested prompts varied in complexity, including simple requests such as *Create a math problem that uses {S}* to ones that include details around their students’ needs, e.g. *This problem will need to be simplified to support diverse learners in the resource room*. The diversity within teachers’ sug-

gested prompts, though, could be informative for future work that extends the evaluation of problem generation beyond template-based inputs.

We observed that some teachers wished to teach multiple standards within a single generated problem, but to scope this study, we simplify all prompt templates to include only a single standard. Thus, to make our setup reflective of standards that may be realistically taught in isolation within a problem, when sampling standards to insert into prompts, we chose ones that appear in MathFish 🐟 with singly-labeled problems.

## E.2 Other Observations of Generated Problems

Aside from standards alignment issues, teachers wrote down additional observations that pertain to the quality of generated problems. We outline two common ones here, in case they are informative for guiding more extensive future work.

**Readability.** Teachers commented on the readability of generated problems. Our prompts do not explicitly indicate the audience of these generations, and generally, these problems could serve two overarching purposes: materials for teachers to work through with students, or materials placed directly in front of students to work on independently. For the former case, some teachers, especially those annotating high school problems, noted that LaTeX formatting outputted by models was difficult for them to parse. For the latter case, some generated problems, especially those addressing lower grades' standards, were not suitable for those students' reading levels.

**Cultural competency.** Some teachers also commented that LMs' attempt to produce culturally competent or culturally responsive practice items was only done at the surface level, e.g. activities that involve "celebrating diversity," without deeper engagement with established frameworks around culturally responsive pedagogy (e.g. [Bonner, 2021](#); [Ladson-Billings, 2021](#)). In addition, it's possible that teachers' prompts were underspecified, and to produce actually culturally aligned problems, teachers needed to explain the context in which they are teaching. For example, one teacher wrote, *culturally responsive leads to an African village? Doesn't seem truly culturally relevant to most students in the US*. Local context also matters, as some topics are prohibited in some schools. For example,

one teacher commented that one generated problem mentioning same-sex marriage and non-binary people is not legally allowed to be taught in public schools in their state.

Finally, we also observe cases where generations from different models contain eerily similar wording. For example, Llama-2 and Mixtral both generated problems containing *A bag contains 5 red balls, 7 blue balls, and 3 green balls*. This can suggest memorization of pretraining or finetuning data, and implies a lack of linguistic or topical diversity among problems generated across models.

## Domain descriptions

- Counting & Cardinality: students learn to know number names and the count sequence, count to tell the number of objects, and compare numbers.
- Operations & Algebra: students learn to solve problems using algebraic thinking and operations such as addition, subtraction, multiplication, and division. They may learn to identify and explain arithmetic patterns, evaluate or manipulate numerical expressions, and reason with equations and inequalities.
- Number & Operations in Base Ten: students learn to work with the base-ten system and build place value understanding.
- Measurement & Data: students learn to work with data and measure attributes such as time, money, length, area, and volume. They may learn to compare measurements to operations and convert between different units of measure.
- Geometry: students learn to classify geometric figures by their properties, and understand and compare the relationships between them. Students develop and use formulas to compute lengths, areas, and volumes, and they use transformations to generate new shapes and compare existing ones. Geometry can be studied with and without coordinates.
- Number & Operations - Fractions: students learn to understand fractions as numbers and may work with them using addition, subtraction, multiplication, and division.
- Ratios & Proportional Relationships: students learn to recognize, describe, represent, and reason with ratios, rates, proportional relationships, and percent.
- Number Systems and Quantity: students learn to understand the complex number system and reason quantitatively.
- Statistics & Probability: students learn to analyze and produce data distributions, and build understanding of univariate and bivariate data. They learn to interpret data, make statistical inferences, justify conclusions, understand rules of probability, and use probability to make decisions.
- Functions: students learn to define, use, and evaluate functions to model relationships between quantities.
- Modeling: students learn to choose and use appropriate mathematics and statistics to analyze real-world empirical situations, improve decisions, and report on their conclusions and the reasoning behind them.

Figure 9: Brief descriptions we wrote to include as options for  $\mathcal{D}$ -level tagging prompts.

## Prompt 1 for verification task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a description of math content. Your task is to assess whether the problem or activity aligns with the provided description. Answer 'yes' if it does align, and 'no' if it does not.

Example {i}:  
Problem/activity:

{example problem activity}

Description:

{example standard description}

Answer:

{example answer} {example thought}

Now, assess whether the following problem or activity aligns with the provided description.

Problem/activity:

{problem activity}

Description:

{standard description}

Figure 10: Prompt 1 for verification.

### Prompt 4 for verification task

You are a math instructor reviewing problems and activities meant to support K-12 students in learning mathematical skills and concepts. You will be shown a problem or activity obtained from school curriculum and a description of mathematical concepts and skills. Your task is to determine whether the problem or activity can enable students to understand the concepts or skills in the provided description. Answer 'yes' if it does, and 'no' if it does not.

Example {i }:

Concept/skill:

{example standard description}

Problem/activity:

{example problem activity}

Answer:

{example answer} {example thought}

Now, determine whether the following problem or activity can enable students to understand the concepts or skills in the provided description.

Concept/skill:

{standard description}

Problem/activity:

{problem activity}

Figure 11: Prompt 4 for verification.

### Prompt 5 for verification task

You are a skilled math instructor and curriculum specialist for K-12 mathematics, specifically the Common Core. Your job is to assess whether a problem or activity is 'aligned' with a given Common Core standard in Mathematics. A problem is 'aligned' with a standard if the problem or activity helps students fully understand or learn the concept or skill described in the standard. If the problem or activity only helps students learn part but not all of a standard, then it does not align. Answer 'yes' if the problem aligns with the standard or 'no' if not.

Example {i}:

Concept/skill:

{example standard description}

Problem/activity:

{example problem activity}

Answer:

{example answer} {example thought}

Now, assess whether the following problem or activity is 'aligned' with the given Common Core standard in Mathematics.

Concept/skill:

{standard description}

Problem/activity:

{problem activity}

Figure 12: Prompt 5 for verification.

### Prompt 7 for verification task

You are a skilled math instructor and curriculum specialist for K-12 mathematics, specifically the Common Core. Your job is to assess whether a problem or activity is 'aligned' with a given Common Core standard in Mathematics. A problem is 'aligned' with a standard if the problem or activity helps students fully understand or learn the concept or skill described in the standard. If the problem or activity only helps students learn part but not all of a standard, then it does not align. Answer 'yes' if the problem aligns with the standard or 'no' if not.

Example {i}:

Problem/activity:

{example problem activity}

Concept/skill:

{example standard description}

Answer:

{example answer} {example thought}

Now, assess whether the following problem or activity is 'aligned' with the given Common Core standard in Mathematics.

Problem/activity:

{problem activity}

Concept/skill:

{standard description}

Figure 13: Prompt 7 for verification.

### Prompt 10 for verification task

You are a math expert reviewing K-12 curriculum. Does this problem or activity enable students to completely learn the following concept or skill? Answer 'yes' if it does, and 'no' if it does not. Answer no if the problem helps students understand part but not all of a concept or skill.

Example {i}:

Problem/activity:

{example problem activity}

Concept/skill:

{example standard description}

Answer:

{example answer} {example thought}

Now, does the following problem or activity enable students to learn the full intent of the following concept or skill?

Problem/activity:

{problem activity}

Concept/skill:

{standard description}

Figure 14: Prompt 10 for verification.

### Prompt 15 for verification task

You are a math expert reviewing K-12 curriculum to assess whether it addresses specific mathematical standards. Does the problem or activity shown below enable students to learn the full intent of the following concept or skill? Answer 'yes' if it does, and 'no' if it does not.

Example {i}:

Problem/activity:

{example problem activity}

Concept/skill:

{example standard description}

Answer:

{example answer} {example thought}

Now, does the problem or activity shown below enable students to learn the full intent of the following concept or skill?

Problem/activity:

{problem activity}

Concept/skill:

{standard description}

Figure 15: Prompt 15 for verification.

### Prompt 1 for tagging task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}, and format your output as a comma-separated list of options e.g. "A, B, C". {relation definition} Output "none" if none of the {level} below are relevant. DO NOT make up additional {level}.

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response:

{example response}

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Figure 16: Prompt 1 for tagging.

### Prompt 3 for tagging task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

You should first write a paragraph explaining which {level} the problem/activity {relation}, and then output a comma-separated list of options. Respond “none” if the problem/activity {relation} none of the provided {level}, and do not make up additional {level}. Please format your response in two lines, as shown in the example below:

Thought: <your paragraph goes here>

Answer: A, C, E

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Thought: {example thought}

Answer: {example response}

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Here is the problem/activity:

{problem activity}

{Level} options:

{options}

Please output both your thoughts about what {level} this problem {relation}, as well as a comma-separated list of {level}:

Figure 17: Prompt 3 for tagging.



### Prompt 4 for tagging task

You are a math instructor reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

Your response should first begin with a paragraph explaining which {level} the problem {relation}, and then output a comma-separated list of options. Respond "none" if the problem/activity {relation} none of the provided {level}. Do not make up additional {level}. Please format your response in two lines, as shown in the example below:

Thought: <your paragraph goes here>

Answer: A, C, E

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response:

Thought: {example thought}

Answer: {example response}

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Your response:

Figure 18: Prompt 4 for tagging.

### Prompt 5 for tagging task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

Your response should be a 'json' object with two fields:

```
{
  "explanation": your justification for your answer, "answer": a succinct comma-separated list
of option letters, e.g. "A, B, C"
}
```

If the problem/activity {relation} none of the provided {level}, your answer should be "none". Do not make up additional {level}.

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response:

```
{
  "explanation": "{example thought}",
  "answer": "{example response}"
}
```

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Your response:

Figure 19: Prompt 5 for tagging. Note that the brackets for json formatting serve a different function than the brackets indicating slots in which we input problems/activities, options, and other level-specific information.

### Prompt 6 for tagging task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

Your response should be a 'json' object with two fields:

```
{
  "explanation": your reasoning for your answer,
  "answer": a succinct comma-separated list of option letters e.g. "A, B, C"
}
```

For example, if the problem or activity {relation} both options D and E, the "answer" key would map to "D, E".

As another example, if the problem or activity only {relation} option A, the "answer" key would map to "A".

If the problem/activity {relation} none of the provided {level}, the "answer" key would map to "none".

Do not make up additional {level}.

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response:

```
{
  "explanation": "{example thought}",
  "answer": "{example response}"
}
```

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Your response:

Figure 20: Prompt 6 for tagging. Note that the brackets for json formatting serve a different function than the brackets indicating slots in which we input problems/activities, options, and other level-specific information.

### Prompt 10 for tagging task

You are reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

Begin your response with a paragraph explaining which {level} the problem {relation}. You may walk through the act of solving or doing the problem/activity, if possible, to illustrate how it {relation} specific {level}. Then, conclude with a comma-separated list of options as your answer. Respond "none" if the problem/activity {relation} none of the provided {level}, and do not make up additional {level}. Please format your response in two lines, as shown in the example below:

Thought: <your paragraph goes here>

Answer: A, C, E

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response:

Thought: {example thought}

Answer: {example response}

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Your response:

Figure 21: Prompt 10 for tagging.

### Prompt 11 for tagging task

You are a math expert reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

You should first write a paragraph explaining which {level} the problem/activity {relation}, and then output a comma-separated list of options. Respond "none" if the problem/activity {relation} none of the provided {level}, and do not make up additional {level}. Please format your response in two lines. You may walk through how a student may solve or do the problem/activity, if possible, to illustrate how it {relation} one or more {level}.

For example, if the problem or activity {relation} {level} D and E, your response would be:  
Thought: <your paragraph goes here>  
Answer: D, E

As another example, if the problem or activity only {relation} {level} A, your response would be:  
Thought: <your paragraph goes here>  
Answer: A

Example {i}:  
Problem/activity:  
{example problem activity}

{Level} options:  
{example options}

Your response:  
Thought: {example thought} Answer: {example response}  
Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:  
{problem activity}

{Level} options:  
{options}

Your response:

Figure 22: Prompt 11 for tagging.

### Prompt 12 for tagging task

You are a math instructor reviewing K-12 curricular materials. You will be shown a problem or activity obtained from school curriculum and a list of one or more {level}. Your task is to assign the problem or activity to one or more relevant {level} it {relation}. {relation definition}

Your response should be a 'json' object with two fields: "explanation", which includes your reasoning, and "answer", which is a comma-separated list of option letters. For example:

```
{
  "explanation": <your reasoning goes here>,
  "answer": "A, B, C"
}
```

If the problem/activity {relation} none of the provided {level}, your answer should be "none". To help you justify your answer, you may try solving the problem or doing the activity, if possible.

Do not make up additional {level}.

Example {i}:

Problem/activity:

{example problem activity}

{Level} options:

{example options}

Your response, in json format:

```
{
  "explanation": "{example thought}",
  "answer": "{example response}"
}
```

Now, assign the following problem or activity to one or more relevant {level} it {relation}.

Problem/activity:

{problem activity}

{Level} options:

{options}

Your response, in json format:

Figure 23: Prompt 12 for tagging. Note that the brackets for json formatting serve a different function than the brackets indicating slots in which we input problems/activities, options, and other level-specific information.