EDUMATH: Generating Standards-aligned Educational Math Word Problems

Anonymous ACL submission

Abstract

Math word problems (MWPs) are critical K-12 educational tools, and customizing them to students' interests and ability levels can increase learning outcomes. However, teachers struggle to find time to customize MWPs for each student given large class sizes and increasing burnout. We propose that LLMs can support math education by generating MWPs customized to student interests and math education standards. To this end, we use a joint human expert-LLM judge approach to evaluate over 11,000 MWPs generated by open and closed LLMs and develop the first teacher-annotated dataset for standards-aligned educational MWP generation. We show the value of our data by using it to train a 12B open model that matches the performance of larger and more capable open models. We also use our teacher-annotated data to train a text classifier that enables a 30B open LLM to outperform existing closed baselines without any training. Next, we show our models' MWPs are more similar to human-written MWPs than those from existing models. We conclude by conducting the first study of customized LLM-generated MWPs with grade school students, finding they perform similarly on our models' MWPs relative to human-written MWPs but consistently prefer our customized MWPs.[1]

1 Introduction

Math word problems (MWPs) are natural language math questions paired with numerical solutions, and are important elements of K-12 math education because they are an effective tool to verify mastery of a specific math concept (Daroczy et al., 2015; Pearce et al., 2013; Schwartz, 2023; Verschaffel et al., 2020). In all K-12 grades, personalizing MWPs to students' interests and ability levels can support learning outcomes by increasing their interest in math and ensuring problems are within their readiness level (Baker et al., 2020; Bernacki and Walkington, 2018; Nasir and Hand, 2008; Pinkard et al., 2020; Walkington, 2013). Typically, teachers manually write or curate MWPs customized to students' interests or ability levels. However, with many competing responsibilities, limited time, and increasing burnout (Kuncl and Christensen, 2024), teachers must often rely on standard question sets from educational websites or practice tests for end-of-year examinations, which are limited in quantity, not customized, and not always easily searchable for math standards/topics (VDOE, b). We propose that large language models (LLMs) can address this challenge by automatically generating MWPs aligned with students' interests and ability levels given their vast natural language ability and increasing abilities in math reasoning (Ahn et al., 2024).

However, given recent work has found LLMs struggle to generate educational MWPs (Ariyarathne et al., 2025; Christ et al., 2024), it is important to conduct further study before directly using them in educational settings. To this end, we aim to explore and enhance LLMs' capacity to generate educational MWPs tailored to students' interests and ability levels as indicated by alignment with math standards from the United States. Recent work has explored generating educational MWPs with LLMs customized to students' interests (Christ et al., 2024), but not math standards. Other work has explored generating math practice problems, including MWPs, with LLMs that are aligned with math topics (Ariyarathne et al., 2025; Sun et al., 2025), but they only generate problems without associated solutions. Two other works explore generating MWPs with LLMs using reference problems (Niyarepola et al., 2022; Zong and Krishnamachari, 2023), but they too only generate problems without solutions. To be useful to students and teachers, generated problems must have associated solutions written in clear natural language and be aligned with particular math standards (e.g.,

[1] We will release our data, experimental code, and models upon publication.

single-step whole number multiplication not exceeding 100) rather than general math topics like multiplication to allow teachers fine-grained control over generated problems. We refer to this task as standards-aligned educational MWP generation.

To evaluate LLMs' ability to generate standards-aligned educational MWPs, we use a joint human expert and LLM judge approach to assess over 11,000 MWPs generated by LLMs based on four criteria that make MWPs educational and standards-aligned. We adapt the first three criteria, *Solvability*, *Accuracy*, and *Educational Appropriateness*, from Christ et al. (2024) and introduce *Standards Alignment*, a new criterion. The nuance needed to simultaneously identify problems that have solutions readable for teachers/students, are educationally appropriate, and effectively incorporate a math standard motivates the need for involving real teachers in evaluations, as it can take teachers many years to develop this expertise.

We conduct this study in five phases. In phase one, we use a joint expert-LLM judge approach to annotate a small existing human-written MWP dataset for grade school math standards given there are no datasets fully annotated for these standards. In phase two, we use the labeled dataset to prompt an existing LLM to generate over 3,000 MWPs that teachers annotate for our four criteria to expand the amount of training and prompting data available for this task. In phase three, we train a mid-size (12B) LLM using annotation data from the second phase to show the value of our data for training standards-aligned educational MWP generators. In this stage, we further show the quality of our annotated data by using it to train a simple text classifier that filters outputs from both our trained model and an existing 30B LLM without further training. In phase four, we evaluate over 8,000 MWPs from our 12B and 30B models as well as several open and closed baselines with an LLM judge aligned with feedback from teachers. In the final phase, we conduct the first evaluation of customized LLM-generated MWPs with real grade school students.

Through these evaluations, we find a performance gap in the task of generating standards-aligned educational MWPs between open and closed models and between smaller and larger open models. We show the value of our data by demonstrating both that training on it is sufficient to nearly close the gap in performance between smaller and larger open LLMs and filtering outputs using our trained text classifier is enough to help a 30B LLM outperform closed baselines without any training. We also show that our data and models' outputs are higher quality and more similar to existing human-written data than outputs from comparison models on a wide range of automatic metrics. Finally, we show that students prefer our generated MWPs to those that are human-written while performing similarly on them. Our key contributions are:

- We find existing open models, especially smaller ones, struggle at the task of generating standards-aligned educational MWPs relative to closed baselines.
- We annotate over 3,000 MWPs with real teachers, which we filter to create the first training dataset for standards-aligned educational MWP generation, the Standards-Targeted Educational Math dataset (STEM).
- We use our data to create two models for standards-aligned educational MWP generation that achieve SOTA performance, which we release alongside all of our annotated data.
- We conduct the first study of custom LLM-generated MWPs with grade school students, demonstrating their value in K-12 education.

2 Related Work

Math Question/Answer (QA) Datasets There are several popular datasets for training and evaluating LLMs on grade school math, including GSM8K (Cobbe et al., 2021), NumGLUE (Mishra et al., 2022), GSM-Hard (Gao et al., 2023), ASDIV (Miao et al., 2020) and SVAMP (Patel et al., 2021). Many works have also developed additional grade school math benchmark datasets or large training datasets consisting of different solution rationales for existing datasets, particularly GSM8K, including both those that are human written (Kim et al., 2023; Mishra et al., 2023) and synthetic (Mitra et al., 2024; Shi et al., 2023; Toshniwal et al., 2024; Yu et al., 2023; Yuan et al., 2023; Yue et al., 2023).

Because existing datasets are designed to train and evaluate grade school math reasoning, they are not aligned with training standards-aligned educational MWP generators. Training data for standards-aligned educational MWP generators must contain high-quality grammar, have both questions and answers, contain readable solutions for teachers and students, be segmented by grade level

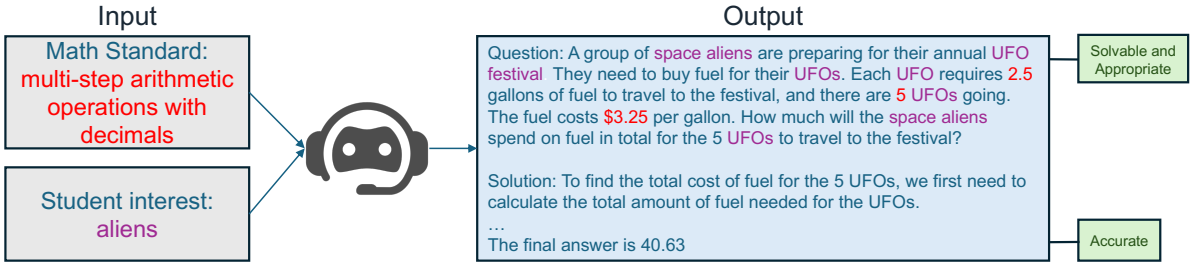

Figure 1: Generating standards-aligned educational MWPs with LLMs. To be educationally useful, problems must be solvable, accurate, educationally appropriate, and aligned with a math standard.

and standards, have questions that are appropriate for the given grade levels and standards, have questions that are similar to those that are written by educators, and contain content appropriate for a school setting. We consider high-quality grammar and similarity to MWPs students would see in a classroom setting as minimally sufficient criteria any potentially useful training dataset must possess. Existing datasets that contain these minimum sufficiency criteria and are aligned with one or more of the other criteria defined above are GSM-Hard (Gao et al., 2023), GSM8K (Cobbe et al., 2021), MathInstruct GSM8K (Yue et al., 2023), ASDIV (Miao et al., 2020), SVAMP (Patel et al., 2021), EGSM (Christ et al., 2024), and Mathwizards (Ariyarathne et al., 2025), although none of these datasets contain all criteria.

Math Word Problem (MWP) Generation Closest to our work is Ariyarathne et al. (2025), who use LLMs to generate elementary school MWPs aligned with a particular math topic like single-digit addition. However, they do not automatically generate solutions, nor do they align their problems with the full language of math standards, which specify both topics and complexity constraints MWPs must follow (e.g., single-step single-digit addition). In addition, they do not conduct evaluations involving real teachers and, instead, rely on automated evaluations and evaluations from non-education college students. Another closely related work is Sun et al. (2025), who use LLMs to generate math problems aligned with a math topic and difficulty level. However, they do not automatically generate solutions, nor are their problems aligned with math standards; instead, they are aligned with general math topics like the Pythagorean theorem. Further, they do not involve teachers in their evaluations and instead rely only on automatic evaluation methods.

Most other MWP generation works use LLMs or traditional NLP approaches to generate MWPs based on pre-specified equations, re-write existing MWPs, or provide additional MWPs based on reference problems (Jiao et al., 2023; Koncel-Kedziorski et al., 2016; Kumar and Sreenivasa Kumar, 2025; Norberg et al., 2023; Niyarepola et al., 2022; Qin et al., 2023, 2024; Wang et al., 2021; Wu et al., 2022; Zhou and Huang, 2019; Zong and Krishnamachari, 2023; Zhou et al., 2023), each of which limit the originality and range of outputs and can often lead to rephrasing training data. These approaches also only generate problems without solutions and require additional input from users, which is infeasible for teachers or students who want to generate customized MWPs, making them incomparable to our work. To address these issues, we generate MWP question/answer pairs based on a grade level and standard simultaneously without requiring input reference problems or equations.

3 Methods

We define math word problems (MWPs) as natural language questions describing contextual math scenarios paired with solutions written in a natural language, chain-of-thought (CoT) (Wei et al., 2023) format. We study generating MWPs with LLMs aligned with a math standard provided in the prompt, as shown in Figure 1. We evaluate generated MWPs based on the grade levels, standards, evaluation criteria, and automatic metrics defined below and conduct a multi-stage pipeline to generate training data and train models for the task of standards-aligned educational MWP generation.

Grade Levels and Standards Studied We study generating MWPs for grades 3-5 in the US. We choose to study grades 3-5 for two main reasons: 1) choosing a narrow grade range allows us to fully assess models' abilities to generate MWPs for all relevant math standards in these grade levels, as each grade has many standards that can be assessed with MWPs, and 2) these are the grade levels for

which MWPs are most used and, therefore, where a standards-aligned educational MWP generator would be most useful (Daroczy et al., 2015; Pearce et al., 2013; Schwartz, 2023; Verschaffel et al., 2020). We use MWP math standards from the Virginia Standards of Learning (VA SOL) because they reference bounded difficulty levels and number ranges for all standards, while the Common Core standards, which are used in many US states, are not always as specific (Initiative; VDOE, a). We provide a mapping of VA SOLs to Common Core standards in the repo associated with this paper.

Evaluation Criteria We evaluate MWPs on four criteria. The first three criteria, *Solvability*, *Accuracy*, and *Educational Appropriateness*, are adapted from Christ et al. (2024). *Solvability* evaluates if questions are possible to solve and have one correct answer. *Accuracy* assesses if generated CoT solutions are accurate. We label MWPs with accurate final answers but inaccurate, confusing, or overly complicated reasoning as inaccurate because our solutions must be readable for teachers and students. *Educational Appropriateness* evaluates if teachers would feel comfortable giving the MWP to one of their students. In general, an educationally appropriate MWP is one that makes sense, does not contain grammar errors or conflicting information, and is about topics that are appropriate for 3rd-5th grade students in a school setting. The final criterion we introduce in this work is *Standards Alignment*, or if the MWP effectively incorporates a pre-specified math standard. Because the latter three criteria are hard to precisely define, we follow Christ et al. (2024) in emphasizing teacher involvement in evaluating MWPs. See Appendix G for MWPs that do and do not meet these criteria.

Automatic Evaluation Metrics Readability automatically assesses the appropriateness of MWPs. We use a composite of 8 readability metrics as implemented by Georgia State University's ARTE tool (GSU) to comprehensively assess readability: the Automated Readability Index (ARI) (Kincaid et al., 1975), Flesch-Kincaid Grade Level (FKGL) (Flesch, 1948), Flesch Reading Ease (FRE) (Kincaid et al., 1975), CAREC and CAREC-M (Crossley et al., 2019), Coh-Metrix L2 Reading Index (CML2RI) (Crossley et al., 2008), New Dale-Chall (Chall and Dale, 1995), and the Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969). We use min-max scaling to normalize each metric in the range of [0, 1] and flip the CML2RI and FRE scores so that higher scores represent more difficult to read texts to match the other metrics. We then report the average of these scaled metrics to capture the overall reading complexity of each MWP. In general, lower reading levels are preferred for MWPs because high reading levels can harm student performance, especially for those struggling in math (Walkington et al., 2018). We also calculate each MWP's average token length, which is a proxy for complexity given that longer questions often include more mathematical operations.

Following prior work (Christ et al., 2024; Jiao et al., 2023; Zhou et al., 2023), we use Perplexity (PPL) and BERTScore to evaluate the quality of our synthetic MWPs and compare them to human-written MWPs, respectively. Lower PPL suggests higher-quality outputs and we calculate PPL with GPT-2. We use BERTScore (Zhang et al., 2020) to compute the semantic similarity of generated MWPs to each other and compare it to that of human-written MWPs. A higher BERTScore for generated MWPs suggests they are more similar to each other than human-written MWPs, and a lower BERTScore suggests the opposite. We also compute BERTScore for generated MWPs relative to human-written MWPs to identify if they are similar and use the BERTScore of human-written MWPs relative to each other as a reference to compare against. For all comparisons to human-written MWPs, we use the 3rd-5th grade subset of ASDIV since it is the only human-written English MWP dataset labeled for grade level.

3.1 Labeling Existing Data for Math Standards

MWPs with labels for math standards are required for training or prompting models to generate standards-aligned educational MWPs. However, no grade school math question/answer datasets are labeled for standards. ASDIV (Miao et al., 2020) has labels for grade and math topics like addition, but not standards, so we used a four-stage approach to label its 3rd-5th grade MWPs (n = 1,027) for VA SOLs. In stage one, an undergraduate education student labeled MWPs for all relevant VA SOLs, as MWPs may address more than one standard. In stage two, a research team member with K-12 teaching experience checked the labels and made adjustments, often reducing the number of VA SOLs associated with MWPs in cases where the student over-labeled MWPs. We also identified and removed two MWPs that were unsolvable and

rewrote several others to make them solvable, as some referenced constraints described in previous MWPs that are not included in the current MWP.

In stage three, we prompted Llama 3.3 70B IT (Meta) with each ASDIV MWP and its VA SOLs from stage two and asked if the question included each VA SOL. See Appendix I.1 for the prompt. We also used a few-shot prompt (see Appendix I.2) for generating readable CoT solutions for the ASDIV subset given it does not have solutions in this format. We compared generated solutions against ground-truth answers to ensure solution accuracy. In stage four, our research team member with K-12 teaching experience reviewed the Llama 3.3 standard labels and made adjustments to remove VA SOLs that were not incorporated in MWPs. To ensure solution quality, we had Gemma 3 27B IT (Team et al., 2025) review generated solutions and rewrite any it deemed were too complex or contained incorrect/unnecessary intermediate reasoning (see Appendix I.3 for the prompt). Finally, the undergraduate education student reviewed solutions to ensure intermediate reasoning steps were accurate and the solution was readable for teachers/students. This resulted in a final dataset with 1,025 MWPs with gold labels for VA SOLs and CoT solutions (see Appendix H for an example).

3.2 Synthetic Data Generation and Expert Annotation

While our labeled dataset provides MWPs for prompting, it does not have enough MWPs for training a standards-aligned educational MWP generator given its small size. As shown in Appendix C, most of the 23 VA SOL combinations in the labeled ASDIV subset also only have a few MWPs they are associated with. Further review of VA SOLs also revealed another 15 combinations of standards that can be addressed by MWPs but are not in the labeled ASDIV subset. To address these gaps, we generated a large synthetic dataset (n = 3,012) and evaluated it with teachers. We generated this data with Llama 3.3 70B IT (Meta) using the prompt in Appendix I.4. We generated roughly 90 samples per VA SOL combination that is not in our labeled ASDIV subset and roughly 80 samples per VA SOL combination that is in the labeled subset but has less than 100 samples to ensure sufficient coverage of all relevant VA SOLs while emphasizing those that are not already included.

Next, teachers annotated our generated MWPs through Prolific. Annotators (n = 1,372) identified as US teachers using Prolific screening criteria and were paid $12/hour. Annotators labeled MWPs for solvability, accuracy, educational appropriateness, and standards alignment using the directions in Appendix A. See Appendix A for annotator demographics and time spent per MWP. Every MWP was annotated twice; if the annotators disagreed on any of the criteria, the MWP was annotated a third time. We assigned final labels based on majority vote. We labeled a question as meeting all criteria (MaC) if it was labeled as solvable, accurate, educationally appropriate, and standards-aligned.

The first two annotators agreed on solvability 90.1 ± 1.2% of the time, accuracy 76.6 ± 1.5% of the time, educational appropriateness 77.5 ± 1.5% of the time, standards alignment 75.3 ± 1.6% of time, and MaC 65.5 ± 1.7% of the time. The agreement rate for solvability and appropriateness is higher than that reported in Christ et al. (2024), while the agreement rate for accuracy and MaC is lower. The lower accuracy agreement rate could be because accuracy in this work evaluates both final answer and intermediate reasoning accuracy as well as solution readability, whereas Christ et al. (2024) only evaluated final answer accuracy. The lower MaC agreement rate could be because it is based on agreement across four metrics instead of three, as was the case in Christ et al. (2024). However, to ensure only the highest quality data are included in our synthetic training dataset, we had each MWP annotated by Gemma 3 27B IT using few-shot examples and the same directions given to annotators (see Appendix I.5 for the prompt). We flipped the label for any MWP humans identified as MaC but Gemma identified as not MaC, but retained the label for any MWP humans identified as not MaC but Gemma identified as MaC.

3.3 Standards-Targeted Educational Math Dataset

Our annotation process identified 1,552 MWPs that MaC, and we compile these MWPs along with the labeled ASDIV subset into a new dataset we call Standards-Targeted Educational Math (STEM). Table 1 shows STEM's advantages over other grade school MWP datasets. Notably, STEM is the only dataset annotated by teachers for math standards and with solutions designed to be readable for students and teachers, making it best suited for training a standards-aligned educational MWP generator. On other metrics, STEM is similar to existing datasets. STEM's question length is in the middle

Dataset	N	Contains Answers	Teacher Annotated	Readable Solutions	Full Math Standard Annotated	Length	Readability	BF1
Mathwizards	3,999	✗	✗	✗	✗	39.5 (14.5)	0.365 (0.179)	73.0
GSM-Hard	1,319	✓	✗	✗	✗	72.9 (25.6)	0.367 (0.178)	73.1
GSM8K	8,792	✓	✗	✗	✗	67.0 (24.4)	0.371 (0.177)	73.1
MathInstruct GSM8K	6,403	✓	✗	✗	✗	66.2 (23.9)	0.371 (0.178)	73.2
ASDIV	2,305	✓	✗	✗	✗	45.1 (15.8)	0.363 (0.179)	73.6
SVAMP	1,000	✓	✗	✗	✗	47.3 (11.7)	0.356 (0.178)	75.9
EGSM	2,093	✓	✓	✗	✗	57.2 (15.7)	0.358 (0.164)	73.1
STEM (Ours)	2,577	✓	✓	✓	✓	53.9 (18.4)	0.384 (0.177)	73.6

Table 1: Characteristics of datasets that can be used to train grade school MWP generators. N is the deduplicated number of questions, Length is average question length (in tokens), readability is measured by a composite of 8 readability metrics, and BF1 is BERTScore F1. Standard deviations, where applicable, are in parentheses.

of the range for grade school math datasets, while its average readability is slightly higher than these datasets, suggesting it may be a little more difficult to read. STEM's BERTScore is similar to that of human-written datasets, suggesting its MWPs are similar. Overall, these metrics highlight that STEM is similar to human-written data but tailored to standards-aligned educational MWP generation.

3.4 Finetuning on Annotated Data

To explore training a standards-aligned educational MWP generator and validate the quality of our labeled data, we use it to finetune a mid-size open LLM, Gemma 3 12B IT (Team et al., 2025). We first conduct Supervised Finetuning (SFT) using STEM. Next, we conduct Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) on our SFT model with our annotated dataset given MaC labels are binary. Next, we train a ModernBERT text classifier (Warner et al., 2024) using our annotation data to label outputs for MaC that reaches 79% accuracy and a 0.861 AUC-ROC on a held-out test set. We then use this classifier on top of our trained KTO model to filter outputs to only those labeled as high quality, resulting in the model we call EDUMATH 12B. In Section 4, we also stack our classifier on top of Qwen 3 30B (Yang et al., 2025) to create EDUMATH 30B. Appendix B.3 shows ablations for EDUMATH 12B's training using our Gemma 3 27B IT annotator. Both finetuning stages and automatic filtering using our text classifier significantly improve model performance, showing the quality of our data. See Appendix B for further training details and hyperparameters.

4 Evaluating EDUMATH Models

Next, we rigorously evaluate both EDUMATH models' generated MWPs using our evaluation criteria and automatic evaluations. We compare 1,000 MWPs from each EDUMATH model to those of open and closed models. For closed models, we evaluate 250 MWPs from each of GPT-4o (OpenAI et al., 2024), GPT-4.1 (OpenAI, 2025), and GPT-4.5 (OpenAI, 2025). We use a smaller set of MWPs for closed models due to API cost. For open models, we evaluate 1,000 MWPs from Gemma 3 12B IT (Team et al., 2025), Gemma 3 27B IT (Team et al., 2024), Qwen 3 30B IT (Yang et al., 2025), and Qwen 3 235B IT (Yang et al., 2025), which all represent highly capable open baselines of various sizes. We prompt each model using 8-shot examples from STEM for a given math standard and ask it to create a new MWP based on the same math standard (see Appendix I.4 for prompting details).

4.1 Main Results

Performance on Evaluation Criteria Using our Gemma 3 27B IT annotator, we evaluated each model's MWPs for solvability, accuracy, educational appropriateness, and standards alignment. To reduce LLM calls, we annotate for all criteria simultaneously and have the LLM output if the MWP MaC along with a CoT explaining its reasoning (see Appendix I.5 for an example). To validate our labels, our research team member with K-12 teaching experience and undergraduate education student labeled 100 randomly selected samples from our evaluation set. The annotators agreed with each other 76% of the time on MaC and had an average agreement rate of 75% with Gemma 3 27B IT for MaC. Given the agreement rates are nearly identical and higher than those reported among teachers in Section 3.2, we are confident in our labels.

Table 2 shows the MaC rate for each model. Our 30B model outperforms all open and closed baselines in MaC, while our 12B model outperforms Gemma 3 27B IT and nearly matches the perfor-

	Model	PPL ↓	BF1	ASDIV BF1	Question Length	Solution Length	Question Readability	Solution Readability	MaC
API	GPT-4o	16.1 (0.33)	75.2	74.2	63.5 (1.10)	152.3 (5.13)	0.378 (0.004)	0.399 (0.004)	92.8 (1.6)
	GPT-4.1	16.3 (0.35)	75.3	74.3	64.5 (1.3)	150.2 (4.34)	0.385 (0.004)	0.400 (0.004)	92.8 (1.6)
	GPT-4.5	15.7 (0.33)	75.0	74.1	61.8 (1.09)	150.6 (4.06)	0.383 (0.004)	0.399 (0.004)	92.0 (1.7)
Public	Gemma 3 12B IT	11.3 (0.10)	77.2	74.0	84.7 (0.72)	240.8 (3.49)	0.372 (0.002)	0.376 (0.002)	63.9 (1.5)
	Gemma 3 27B IT	12.2 (0.12)	77.1	74.5	76.1 (0.73)	215.1 (3.17)	0.371 (0.002)	0.378 (0.002)	75.4 (1.3)
	Qwen 3 30B IT	12.3 (0.10)	76.9	74.0	68.1 (0.70)	202.9 (3.12)	0.383 (0.002)	0.383 (0.002)	87.3 (1.0)
	Qwen 3 235B IT	12.5 (0.12)	76.1	74.1	66.5 (0.64)	186.5 (2.53)	0.392 (0.002)	0.378 (0.002)	89.0 (1.0)
	EDUMATH 12B (Ours)	**9.5 (0.10)**	74.5	73.8	54.9 (0.54)	166.7 (2.45)	0.389 (0.002)	0.381 (0.002)	85.9 (1.2)
	EDUMATH 30B (Ours)	12.0 (0.13)	76.1	73.8	60.4 (0.71)	163.5 (2.28)	0.380 (0.002)	0.379 (0.003)	**94.6 (0.9)***

Table 2: Comparing LLMs for standards-aligned educational MWP generation. All metrics are averages over 1,000 MWPs/model except for GPT models (they include 250 MWPs due to cost). PPL is perplexity, BF1 is BERTScore F1, ASDIV BF1 compares each model's questions to the ASDIV 3rd-5th grade subset, length is average token length, readability is a composite of 8 metrics, and MaC is meets all criteria. Bold indicates the lowest PPL and best MaC performance and a * indicates the difference between the best open MaC performance and second best open MaC performance is significant at the $p < 0.01$ level. Standard errors, where appropriate, are in parentheses.

mance of Qwen 3 30B IT, two much larger and more capable models. The table also shows performance gaps between open and closed models and between smaller open and larger open models that our data/models eliminate or nearly eliminate. These results suggest our data are sufficient for enabling open models to meet SOTA performance.

Comparison to Human-written MWPs Table 2 also shows PPL, BERTScore, and the average question and solution length and readability for each model's MWPs. Our 12B model has the lowest PPL among all models and our 30B model has the lowest PPL among larger models, suggesting their MWPs are high quality. Our 12B model's BERTScore is most similar to ASDIV's within-dataset BERTScore among all models, while our 30B model ties with Qwen 3 235B IT in having a BERTScore closest to ASDIV's among the open models excluding EDUMATH 12B. The BERTScores of our models' MWPs compared to ASDIV is roughly equivalent to ASDIV's within-dataset BERTScore, suggesting their MWPs are similar in diversity to human-written MWPs, while the MWPs from other models have higher BERTScores for this comparison. Our models' average question lengths are most similar to those of STEM and ASDIV, suggesting their complexity, as measured by this proxy, is similar to teacher-written MWPs. Our models have the shortest solutions among open models, though the closed baselines have the shortest solutions among all models. Shorter solutions are preferable to longer, more verbose ones that may be daunting to read.

For readability, all models output questions with readability similar to STEM except that Qwen 3 235B IT, GPT 4.1, and EDUMATH 12B slightly exceed STEM's readability. For solution readability, all open models' solutions have lower readability than STEM, while closed models exceed STEM's readability, denoting a strength of open models. Overall, Table 2 suggests the EDUMATH models' MWPs are high quality and similar to human-written MWPs with the exception that EDUMATH 12B's MWPs are slightly harder to read.

Performance by Math Topic Although there are 38 math standard combinations, they can be grouped into combinations of one or more of 8 math topics: addition/subtraction, multiplication/-division, fractions, decimals, perimeter/area/volume, patterns, elapsed time, and measurement conversion. Exploring MaC rates by topic can highlight the topics each model struggles with and succeeds at. Table 3 shows this comparison, while Appendix D shows further analysis of common error types for each model. EDUMATH 12B performs best at time problems and struggles the most with fraction problems, while EDUMATH 30B performs best at conversion problems and struggles the most with multiplication/division problems. EDUMATH 12B and the Gemma models have larger gaps in performance between the topic they perform best and worse at than the other models. EDUMATH 30B is the only model for which performance on each topic is above 90%, further highlighting its SOTA performance.

5 Evaluating MWPs with Students

While we show our models' MWPs are high-quality and can be customized to student interests (see Appendix E), it is critical to ensure students perform

	Model	Add/Sub	Mult/Div	Fractions	Decimals	Perimeter/ Area/ Volume	Patterns	Time	Conversion
API	GPT-4o	91.7	91.3	88.9	**96.2**	96.4	**100.0**	92.3	**100.0**
	GPT-4.1	89.4	91.3	88.9	**96.2**	98.2	95.0	92.3	**100.0**
	GPT-4.5	90.2	89.3	82.2	92.3	94.6	**100.0**	**100.0**	**100.0**
Public	Gemma 3 12B IT	63.0	59.2	61.5	62.2	79.4	77.0	67.3	66.7
	Gemma 3 27B IT	71.5	69.9	82.3	77.7	80.5	90.2	85.2	74.5
	Qwen 3 30B IT	82.1	81.9	91.1	87.0	93.6	97.6	93.5	88.7
	Qwen 3 235B IT	84.0	83.6	89.5	91.7	97.3	97.6	94.4	94.4
	EDUMATH 12B (Ours)	88.0	81.0	78.6	93.0	92.9	84.8	98.1	92.5
	EDUMATH 30B (Ours)	**94.3**	**92.0**	**92.2**	95.2	**98.8**	97.5	93.5	**100.0**

Table 3: Meets all criteria rate for each model by topic. Bold indicates the best performance for each topic.

similarly on their MWPs relative to human-written MWPs and to identify if students prefer their customized MWPs to non-customized MWPs. To test this, 3rd-5th graders solved a worksheet with one MWP generated by EDUMATH 30B and aligned with topics they are interested in and one written by humans, either from existing data (K5Learning or the ASDIV 3rd-5th grade subset) or written by our research team member with K-12 teaching experience if a worksheet's math topic was not captured in existing data (see Appendix F.3 for an example). We also asked students which MWP they preferred and why. We tested these worksheets in two experimental settings at two elementary schools (see Appendix F.2 for school demographics). All worksheets were screened by students' teachers to ensure the MWPs were appropriate and reasonable. At the first school, 4th graders (n = 82) solved one worksheet for each of four weeks where both the LLM-generated and human-written MWP were the same for all students. Each MWP was aligned with a standard students were currently learning and the LLM-generated MWP was based on a topic teachers identified as interesting to students. At the second school, 3rd-5th graders (n = 12) receiving services through a math interventionist completed one or two worksheets with the same human-written MWP for every student in each grade level and a customized LLM-generated MWP for each student in that grade level, depending on how many standards they were learning (see Appendix F.2 for details). To inform the customized MWPs, the math interventionist surveyed students to identify topics they were interested in, resulting in a high level of customization for each student. These two experimental conditions allow us to test differential effects of customization at the class and individual level, while also testing our MWPs with students across the grade levels we study.

Through this experiment (see Appendix F.1 for full results), we found students performed similarly on human-written and LLM-generated MWPs, further suggesting EDUMATH 30B's MWPs are human quality. Students preferred the LLM-generated MWPs over the human-written MWPs in both conditions. In School 2 where customization was at the individual level, every student but one preferred the LLM-generated MWPs. Students most often reported liking the LLM-generated MWP because they liked the topic it was about. These results offer compelling evidence that LLMs can support elementary math education by providing customized MWPs students find exciting.

6 Conclusion

We explore standards-aligned educational MWP generation and create the first dataset to train models for this task and the only one verified by teachers to have readable solutions. We also conduct an evaluation of the performance of existing models at this task and fully release our annotations to aid the research community. We demonstrate the quality of STEM and our annotated data by using them to develop the EDUMATH models, the first standards-aligned educational MWP generators. Our evaluations show EDUMATH 12B matches the performance of larger and more capable open models, while EDUMATH 30B sets a new SOTA. We also conduct the first study of custom LLM-generated MWPs with grade school students, finding they perform similarly on our MWPs to those that are human-written, while preferring our MWPs. Our findings suggest that LLMs are poised to greatly contribute to K-12 education by directly generating customized practice problems for students to reduce teacher burdens. Future research should develop models and data to generate problems for other grade levels and types of math problems.

Limitations

One limitation of this work is that we restricted it to the study of 3rd-5th grade MWPs. While these are the grade levels for which MWPs are most highly used, all grade levels use MWPs to some degree, and we believe other grade levels are therefore compelling areas for future work. Another limitation of our work is that our MWPs are text-only, and many MWPs students encounter are multi-modal, containing both images/tables/figures and text. Generating these multi-modal MWPs is an interesting and challenging direction for future work. While we use a standard prompt for all evaluated models to make fair comparisons across models, future work could examine prompt engineering approaches to improve the performance of specific models. The high cost of human annotation is a limitation of MWP generator studies broadly, and we hope our annotations, trained text classifier, and Gemma 3 27B IT automated annotator can help motivate future work in automatic classification efforts. Lastly, while we aligned our Gemma 3 27B IT automated annotator with human experts over multiple rounds of iteration and reached an average agreement rate on par with that of the agreement rate between human experts, it is possible the model missed nuances in MWPs that only teachers with years of experience can identify. Future work should thus continue to involve teachers in the evaluation process of model outputs.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. *arXiv preprint*. ArXiv:2402.00157 [cs].

Nimesh Ariyarathne, Harshani Bandara, Yasith Heshan, Omega Gamage, Surangika Ranathunga, Dilan Nayanajith, Yutharsan Sivapalan, Gayathri Lihinikaduarachchi, Tharoosha Vihidun, Meenambika Chandirakumar, Sanujen Premakumar, and Sanjula Gathsara. 2025. Elementary Math Word Problem Generation using Large Language Models. *arXiv preprint*. ArXiv:2506.05950 [cs].

Ryan Baker, Wei Ma, Yuxin Zhao, Shengni Wang, and Zhenjun Ma. 2020. The Results of Implementing Zone of Proximal Development on Learning Outcomes. Technical report, International Educational Data Mining Society. ERIC Number: ED608058.

Matthew L. Bernacki and Candace Walkington. 2018. The role of situational interest in personalized learning. *Journal of Educational Psychology*, 110(6):864–881. Place: US Publisher: American Psychological Association.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Google-Books-ID: 2nbuAAAAMAAJ.

Bryan R Christ, Jonathan Kropko, and Thomas Hartvigsen. 2024. MATHWELL: Generating Educational Math Word Problems Using Teacher Annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11914–11938, Miami, Florida, USA. Association for Computational Linguistics.

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. *arXiv preprint*. ArXiv:2110.14168 [cs].

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNAMARA. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly*, 42(3):475–493. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1545-7249.2008.tb00142.x.

Scott A. Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading*, 42(3-4):541–561. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9817.12283.

Gabriella Daroczy, Magdalena Wolska, Walt Detmar Meurers, and Hans-Christoph Nuerk. 2015. Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, 6:348.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. *arXiv preprint*. ArXiv:2402.01306 [cs].

R. Flesch. 1948. A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221–233.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided Language Models. *arXiv preprint*. ArXiv:2211.10435 [cs].

GSU. ARTE.

Common Core State Standards Initiative. Common Core State Standards Initiative – Preparing America's Students for College & Career.

Ying Jiao, Kumar Shridhar, Peng Cui, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. Automatic Educational Question Generation with Difficulty Level Controls. In *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 476–488, Cham. Springer Nature Switzerland.

K5Learning. Free worksheets.

Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak, and Jongwuk Lee. 2023. It ain't over: A multi-aspect diverse math word problem dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14984–15011, Singapore. Association for Computational Linguistics.

J. P. Kincaid, Jr. Fishburne, Rogers Robert P., Chissom Richard L., and Brad S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Rik Koncel-Kedziorski, Ioannis Konstas, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2016. A Theme-Rewriting Approach for Generating Algebra Word Problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1617–1628, Austin, Texas. Association for Computational Linguistics.

Suresh Kumar and P Sreenivasa Kumar. 2025. Validity Checking and Repairing of Machine Generated Transfer-Type Word Problems. *Applied Ontology*, page 15705838241303829. Publisher: SAGE Publications.

Liz Kuncl and Kyle Christensen. 2024. Overworked and Undervalued: Retaining Top Educators. Section: Education.

G. Harry Mc Laughlin. 1969. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8):639–646. Publisher: [Wiley, International Reading Association].

Meta. Llama 3.3 Model Cards and Prompt formats.

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.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2023. Lila: A Unified Benchmark for Mathematical Reasoning. *arXiv preprint*. ArXiv:2210.17517 [cs].

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks. *arXiv preprint*. ArXiv:2204.05660 [cs].

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *Preprint*, arXiv:2402.14830.

Na'ilah Suad Nasir and Victoria Hand. 2008. From the Court to the Classroom: Opportunities for Engagement, Learning, and Identity in Basketball and Classroom Mathematics. *Journal of the Learning Sciences*, 17(2):143–179. Publisher: Routledge _eprint: https://doi.org/10.1080/10508400801986108.

Kashyapa Niyarepola, Dineth Athapaththu, Savindu Ekanayake, and Surangika Ranathunga. 2022. Math Word Problem Generation with Multilingual Language Models. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 144–155, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Kole Norberg, Husni Almoubayyed, Stephen E Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steve Ritter. 2023. Rewriting math word problems with large language models. *Proceedings of the Workshop on Empowering Education with LLMs-the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023)*, 3487:163–172.

OpenAI. 2025. Introducing GPT-4.5.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? *arXiv preprint*. ArXiv:2103.07191 [cs].

Daniel Pearce, Faye Bruun, Kim Skinner, and Claricia Lopez-Mohler. 2013. What teachers say about student difficulties solving mathematical word problems in grades 2-5. *International Electronic Journal of Mathematics Education*, 8:3–19.

Nichole Pinkard, C. K. Martin, and S. Erete. 2020. Equitable approaches: opportunities for computational thinking with emphasis on creative production and connections to community. *Interactive Learning Environments*, 28(3):347–361. Publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2019.1636070.

Longhu Qin, Jiayu Liu, Zhenya Huang, Kai Zhang, Qi Liu, Binbin Jin, and Enhong Chen. 2023. A Mathematical Word Problem Generator with Structure Planning and Knowledge Enhancement. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1750–1754, New York, NY, USA. Association for Computing Machinery.

Wei Qin, Xiaowei Wang, Zhenzhen Hu, Lei Wang, Yunshi Lan, and Richang Hong. 2024. Math Word Problem Generation via Disentangled Memory Retrieval. *ACM Transactions on Knowledge Discovery from Data*. Just Accepted.

Sarah Schwartz. 2023. Why Word Problems Are Such a Struggle for Students—And What Teachers Can Do. *Education Week*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Jianwen Sun, Wangzi Shi, Xiaoxuan Shen, Shengyingjie Liu, Luona Wei, and Qian Wan. 2025. Multi-objective math problem generation using large language model through an adaptive multi-level retrieval augmentation framework. *Information Fusion*, 119:103037.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 178 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*. ArXiv:2408.00118 [cs].

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv: Arxiv-2402.10176*.

VDOE. a. K-12 Standards & Instruction.

VDOE. b. SOL Practice Items (All Subjects).

Lieven Verschaffel, Stanislaw Schukajlow, Jon Star, and Wim Van Dooren. 2020. Word Problems in Mathematics Education: A Survey. *ZDM: The International Journal on Mathematics Education*, 52(1):1–16. Publisher: Springer ERIC Number: EJ1243930.

Candace Walkington, Virginia Clinton, and Pooja Shivraj. 2018. How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal*, 55(2):362–414. Place: US Publisher: Sage Publications.

Candace A. Walkington. 2013. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105(4):932–945. Place: US Publisher: American Psychological Association.

Zichao Wang, Andrew S. Lan, and Richard G. Baraniuk. 2021. Math Word Problem Generation with Mathematical Consistency and Problem Context Constraints. *arXiv preprint*. ArXiv:2109.04546 [cs].

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint*. ArXiv:2412.13663 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Qinzhuo Wu, Qi Zhang, and Xuanjing Huang. 2022. Automatic Math Word Problem Generation With Topic-Expression Co-Attention Mechanism and Reinforcement Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1061–1072.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *Preprint*, arXiv:2309.12284.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models. *arXiv preprint*. ArXiv:2308.01825 [cs].

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning. *arXiv preprint*. ArXiv:2309.05653 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Variable	Mean
Age	39.3 (11.7)
Time Taken Per MWP (in seconds)	117.0 (75.0)
	Percentage
Female	67.1
Race or Ethnicity	
White	73.6
Black	12.0
Asian	5.9
Mixed	5.4
Other	2.8
Prefer not to say	0.3

Table 4: Annotator demographics and average time spent per MWP. Standard deviations, where appropriate, are in parentheses.

Qingyu Zhou and Danqing Huang. 2019. Towards Generating Math Word Problems from Equations and Topics. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 494–503, Tokyo, Japan. Association for Computational Linguistics.

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. *Preprint*, arXiv:2306.09064.

Mingyu Zong and Bhaskar Krishnamachari. 2023. Solving Math Word Problems concerning Systems of Equations with GPT-3. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15972–15979. Number: 13.

A Annotator Directions and Demographics

Figures 2, 3, 4, and 5 display the directions presented to annotators for annotating MWPs for solvability, accuracy, educational appropriateness, and standards alignment, respectively. If an annotator answered "No" to solvability or any other criteria, they were instructed to select "NA" for the remaining criteria given that a failure to meet one of the criteria automatically means a MWP does not MaC. This saved the annotators time, as they did not need to unnecessarily annotate the remaining criteria. Table 4 displays annotator demographics and average time spent per MWP. Our human annotation study was approved by the institutional review board at our host university.

B Model Training Details and Hyperparameters

B.1 EDUMATH 12B

For the supervised finetuning (SFT) training stage, we trained Gemma 3 12B IT on the STEM dataset for 5 epochs using 4 A6000 GPUs, a 85/15 train/-validation split, a learning rate of 1e-6, a batch size of 1, and a 10% warm-up ratio. We saved the model every 1,500 training steps and selected the model with the lowest validation loss as our final model (this was the model after step 10,000). For the Kahneman-Tversky Optimization (KTO) training stage, we continued training our SFT model on a dataset of 4,039 rows that combined our annotated data and the ASDIV 3rd-5th grade subset. We trained on this dataset using 4 A100 80GB GPUs for 5 epochs with an 85/15 train/validation split, learning rate of 5e-6, batch size of 8, a 10% warm-up ratio, and desirable/undesirable weights based on inverse class frequency. We saved the model every 500 steps and selected the model with the lowest validation loss as our final model, which was the final model checkpoint.

B.2 Text Classifier

We trained our ModernBERT (Warner et al., 2024) text classifier such that a prediction of 1 means a MWP does not MaC and a prediction of 0 means a MWP does MaC. The training dataset combined our annotation data and MWPs from the ASDIV 3rd-5th grade subset that addressed standards that were not already in our annotated dataset for a total of 3,664 rows. We combined the data in this way to ensure that all standards were represented in our training data without over-representing positive examples for standards that were already in the annotated dataset. Given that there were a higher share of MWPs that MaC after adding in the ASDIV 3rd-5th grade subset (39.9% of rows did not MaC), we trained the model using a loss term weighted by the inverse class balance. We trained the model using an A6000 GPU for 10 epochs with an 80/10/10 training/validation/test split, learning rate of 1e-5, batch size of 8, and a 10% warm-up ratio, selecting the model checkpoint with the lowest validation loss as our final model (this was the model checkpoint after epoch 6). This model achieved a test accuracy of 79.0% and AUC-ROC of .861.

Is this question solvable? A solvable question means that it can be solved with the information present and does not contain a mathematical scenario that is impossible (e.g., giving away more money or items than you have).

- Yes

- No

Figure 2: Annotator directions and response options for annotating MWPs for solvability.

Is the solution for this question correct? If the answer is correct but the reasoning is wrong, does not make sense, and/or is too complicated for a student/teacher to read, answer "No."

- Yes, the solution and reasoning are correct.

- No. The solution is incorrect.

- No. The solution is correct, but the reasoning contains an error, does not make sense, and/or is too complicated for a student/teacher to read.

- NA

Figure 3: Annotator directions and response options for annotating MWPs for accuracy.

Would you feel comfortable giving this question to a student in a 3rd-5th grade school setting?

- Yes: This question is understandable and appropriate for a school setting.

- No: This question contains material inappropriate for a school setting (e.g., language about harming someone).

- No: This question is strange, confusing, contains conflicting information, and/or is not based in reality (e.g., contains misinformation).

- No: This question requires no mathematical operations to solve because it gives the answer away.

- No: It is inappropriate for a different reason.

- NA

Figure 4: Annotator directions and response options for annotating MWPs for educational appropriateness.

Does the question adequately address important elements from the pre-specified math topic(s)?

If more than one math topic is listed, then the question should incorporate important elements of EACH listed math topic. If one of the topics lists multiple mathematical operations like addition, subtraction, and division, it is okay if the question just addresses one of those operations.

Select the most appropriate option below.

- Yes: The question addresses important elements from the specified topic(s)

- No: The question is too hard for the given topic(s).

- No: The question does not address some important parts of the specified topic(s).

- No: The question does not address the specified topic(s) at all.

- No: The question requires additional math topics or operations that are not listed in the specified math topic(s).

- NA

Figure 5: Annotator directions and response options for annotating MWPs for standards alignment.

Model	MaC
Gemma 3 12B IT	63.9 (1.5)
SFT	76.2 (1.3)*
KTO	81.0 (1.2)*
KTO + Classifier	85.9 (1.2)*

Table 5: EDUMATH finetuning stages ablation. MaC denotes the share of 1,000 MWPs from each model that meet all criteria according to our Gemma 3 27B IT annotator. Standard errors are in parentheses and a * indicates a statistically significant difference between the current and previous stage at the $p < 0.01$ level.

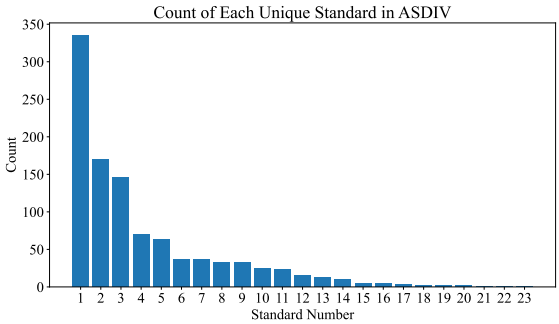

Figure 6: Count for each standard in the ASDIV 3rd-5th grade subset. A mapping of each standard number to its text is provided in the paper repo.

B.3 Finetuning Ablation

Table 5 shows an ablation for each EDUMATH 12B finetuning stage. Notably, each stage results in a statistically significant increase in MaC relative to the preceding stage, demonstrating the value of our annotated data for training standards-aligned educational MWP generators.

C Standard Counts in Labeled ASDIV Subset

Figure 6 displays the count of MWPs from each standard included in the ASDIV 3rd-5th grade subset.

D Error Analysis

Common Error Types by Model Using our Gemma 3 27B IT annotator's CoT, we can also compute the frequencies of the types of errors each model makes to highlight their relative strengths/weaknesses. We can look at the proportion of overall MaC errors based on solvability, accuracy, educational appropriateness, and standards alignment as well as the proportion of each specific error type made for each criteria individually based on the response options given to human annotators reported in Appendix A (with the exception of solvability, which only has one response option). These frequencies are calculated by prompting Gemma 3 27B IT to identify the specific criteria failure and error type for that criteria in the Gemma 3 27B IT annotator's CoT (see Appendix I.6 for the prompt used for this experiment). We show results for this experiment in Tables 6, 7, 8, and 9.

Failures in any of the four criteria are equally harmful to student learning because a failure in any area makes the question inappropriate to give to a student. As shown in Table 6, overall models tend to avoid making solvability and educational appropriateness errors. The majority of errors for each model, therefore, are for accuracy and standards alignment, which are the more nuanced criteria.

As shown in the annotator directions in Appendix A, there are two primary reasons a question would be labeled as inaccurate: 1) The solution is incorrect or 2) The solution is correct but contains a reasoning error, does not make sense, or is too complicated for a student/teacher to read, both of which would be equally harmful to student learning. As shown in Table 7, all models except GPT 4.5 make a majority of their accuracy errors from the latter error type, highlighting that they tend to arrive at the right answer most of the time, even when they make an accuracy error.

As shown in Appendix A, there are four primary reasons a MWP could be labeled as educationally inappropriate, including containing inappropriate material for a school setting, being strange or unrealistic, requiring no mathematical operations to solve, or being inappropriate for another reason. Of these errors, containing inappropriate material for a school setting would provide the most direct harm to young learners. As shown in Table 8, no models make this most harmful error type, which is an improvement over the results reported in Christ et al. (2024) for older models. EDUMATH 30B and the GPT models avoid educational appropriateness errors altogether, whereas the other models tend to make mistakes in outputting MWPs that are strange or unrealistic, with the exception that Gemma 3 27B IT has an equal share of MWPs that require no mathematical operations to solve.

As shown in Appendix A, there are four reasons why a MWP would not be aligned with a standard or standards: 1) It is too hard for the standard(s), 2) It does not address important parts of the standard(s), 3) It does not address the standard(s) at

	Model	Solvability	Accuracy	Educational Appropriateness	Standards Alignment
API	GPT-4o	5.6	50.0	0.0	44.4
	GPT-4.1	0.0	61.1	0.0	38.9
	GPT-4.5	5.0	50.0	0.0	45.0
Public	Gemma 3 12B IT	7.1	43.1	2.5	47.3
	Gemma 3 27B IT	4.4	48.8	0.8	46.0
	Qwen 3 30B IT	1.5	37.2	0.8	60.5
	Qwen 3 235B IT	0.9	38.6	1.7	58.8
	EDUMATH 12B (Ours)	0.8	31.7	5.0	62.5
	EDUMATH 30B (Ours)	5.5	41.7	0.0	52.8

Table 6: Proportion of overall errors from each evaluation criteria by model, or *P(criteria failure | error)*.

	Model	Wrong Answer	Right Answer, Poor or Confusing Reasoning
API	GPT-4o	33.3	66.7
	GPT-4.1	45.5	54.5
	GPT-4.5	50.0	50.0
Public	Gemma 3 12B IT	45.9	54.1
	Gemma 3 27B IT	42.3	57.7
	Qwen 3 30B IT	22.9	77.1
	Qwen 3 235B IT	36.4	63.6
	EDUMATH 12B (Ours)	44.7	55.3
	EDUMATH 30B (Ours)	26.7	73.3

Table 7: Classification of accuracy errors for each model. See Figure 3 for directions presented to our Gemma 3 27B IT automated annotator.

all, or 4) It requires the use of additional math topics beyond those listed in the standard(s), each of which would be equally harmful to student learning. Table 9 shows the majority of standards alignment errors for each model are for MWPs that are missing important parts of the prespecified standard(s). Many models also frequently make MWPs that require additional math topics to solve, while it is less frequent that they make MWPs that are too hard to solve for the given standard(s) or do not address the standard(s) at all. These results suggest that models are generally able to incorporate some important elements from prespecified standard(s) when writing MWPs even when they do not fully align the problems with those standards.

E Customizing MWPs to Student Interests

While Christ et al. (2024) showed it was trivial to generate MWPs aligned with student interests with LLMs, we ran an experiment to verify this for our EDUMATH models. For this experiment, we generated 1,000 MWPs from both EDUMATH models based on a randomly selected topic from a list of 188 topics 3rd-5th grade students might be interested in. We then instructed Gemma 3 27B IT to indicate whether the MWP effectively incorporated the prespecified topic, which we used to calculate the proportion of generated MWPs that successfully included a topic (see Appendix I.7 for details on the prompts and topics used in this experiment). Based on this experiment, 95.5% of EDUMATH 12B and 97.7% of EDUMATH 30B MWPs successfully included a pre-specified topic, indicating both models are effectively able to incorporate student interests.

F Student Study Details

F.1 Student Study Full Results

Table 10 displays the main results from the student study for each experimental condition, which was approved by the institutional review board at our host university. As shown in the table, students in both experimental conditions solved the human-written and LLM-generated MWPs at a similar rate, with the differences in solve rate being statistically insignificant. However, students in both conditions preferred the LLM-generated MWPs to the human-written MWPs, though the preference for the LLM-generated MWPs was much stronger in School 2 (customization at the individual level),

	Model	Strange	Harmful	No Ops	Other
API	GPT-4o	0.0	0.0	0.0	0.0
	GPT-4.1	0.0	0.0	0.0	0.0
	GPT-4.5	0.0	0.0	0.0	0.0
Public	Gemma 3 12B IT	88.9	0.0	11.1	0.0
	Gemma 3 27B IT	50.0	0.0	50.0	0.0
	Qwen 3 30B IT	100.0	0.0	0.0	0.0
	Qwen 3 235B IT	100.0	0.0	0.0	0.0
	EDUMATH 12B (Ours)	100.0	0.0	0.0	0.0
	EDUMATH 30B (Ours)	0.0	0.0	0.0	0.0

Table 8: Classification of appropriateness errors for each model. See Figure 4 for directions presented to our Gemma 3 27B IT automated annotator.

	Model	Too Hard	Missing Important Parts	Does Not Address Standard	Requires Additional Topics
API	GPT-4o	0.0	100.0	0.0	0.0
	GPT-4.1	0.0	71.4	0.0	28.6
	GPT-4.5	0.0	77.8	11.1	11.1
Public	Gemma 3 12B IT	5.8	61.0	3.5	29.7
	Gemma 3 27B IT	7.7	71.6	1.7	19.0
	Qwen 3 30B IT	3.8	79.5	0.0	16.5
	Qwen 3 235B IT	0.0	85.1	0.0	14.9
	EDUMATH 12B (Ours)	2.7	74.7	9.3	13.3
	EDUMATH 30B (Ours)	5.3	78.9	0.0	15.8

Table 9: Classification of standards alignment errors for each model. See Figure 5 for directions presented to our Gemma 3 27B IT automated annotator.

which is expected. In School 2, every student except one preferred the LLM-generated MWP over the human-written one. Table 11 shows a breakdown of the reasons students indicated for preferring one problem over another in each experimental condition. As shown in the table, in School 1 students who preferred the LLM-generated MWP most often preferred it because they liked the topic it was about (e.g., the student interest used to guide the generation), whereas their preference reason was more varied in cases where they preferred the human-written MWP. In School 2, the vast majority of students preferred the LLM-generated MWP because they liked the topic. In this condition, the one student who preferred the human-written MWP over the LLM-generated MWP preferred it because they liked the topic it was about.

F.2 Student Demographics

Table 12 displays the school-level demographics for both schools included in the student study, while Table 13 shows the number of students per grade level and number of worksheets per student in the second student experimental condition (customization at the individual level). We report demographics at the school level to avoid collecting any personally identifying information for individual students. While the first school is considerably more diverse than the second, all students included from the second school were receiving support from a math interventionist, suggesting they were struggling in school.

F.3 Sample Student Worksheet

Figure 7 shows an example student worksheet used for the first week of data collection in the first experimental condition (student interest customization at the class level rather than individual level).

G Sample Annotated Data

G.1 Example that Meets All Criteria (MaC)

Grade Level: 4

Standard: Single-step or multi-step multiplication with whole numbers using a) a two-digit factor and a one-digit factor, b) a three-digit factor and a one-digit factor or c) a two-digit factor and a

Please solve the word problems below and then answer the questions about which word problem you liked best. Please make sure to show your work.

1. Marcus and his friends are starting a recycling project to help the school raise some money for charity. They were tasked to collect different materials and create useful things from those. If Marcus was able to gather 25 milk bottles, and John was able to gather 20 milk bottles, how many milk bottles do they have available for recycling?

2. Superman needs to collect some dog bones for his best friend Krypto. If he found 15 dog bones in the Fortress of Solitude and 25 dog bones at the Daily Planet, how many dog bones does he have for Krypto?

Which word problem did you like best?

Why did you like that word problem the best?

Figure 7: Sample student worksheet for first experimental condition in the student study. The second MWP is the one customized to a student interest at the class level (Superman).

Variable	Human-written MWP Solve Rate	LLM-generated MWP Solve Rate	Preferred LLM-generated MWP
School 1 (customization at class level)	68.2	70.6	64.0*
School 2 (customization at individual level)	73.7	68.4	94.7*

Table 10: Main results for student study. A * denotes the difference in the share of students who preferred the LLM-generated MWP to the human-written MWP is statistically significant at the $p < 0.01$ level.

	Preference Reason	School 1 (customization at class level)	School 2 (customization at individual level)
LLM-generated	Liked Topic	48.5	77.8
	Easy	17.6	16.7
	Challenging	13.9	0.0
	Liked Math Operation	10.3	5.5
	More Realistic	0.0	0.0
	Contained More Context	3.6	0.0
	Other	6.1	0.0
Human-written	Liked Topic	20.4	100.0
	Easy	25.8	0.0
	Challenging	19.4	0.0
	Liked Math Operation	12.9	0.0
	More Realistic	3.2	0.0
	Contained More Context	0.0	0.0
	Other	18.3	0.0

Table 11: Breakdown of reasons indicated by students when explaining why they preferred one MWP over the other for each experimental condition.

two-digit factor

Question: A Pokémon trainer has 14 boxes of Poké Balls to store her Pokémon. Each box can hold 25 Poké Balls. How many Poké Balls can the trainer store in total?

Solution:
To find the total number of Poké Balls the trainer can store, we need to multiply the number of boxes (14) by the number of Poké Balls each box can hold (25).
14 x 25 = 350
This means the trainer can store a total of 350 Poké Balls.
The final answer is 350.

Annotation: This MWP MaC because it is solvable, has an accurate solution, contains content appropriate for a school setting, and is aligned with the pre-specified educational math standard.

G.2 Solvability Negative Example

Question: Maria is baking cookies for her school bake sale. She plans to make batches of cookies with 2 1/6 cups of flour each. If she has 3 1/6 cups of flour, how much more flour does she need to buy to make all the cookie batches?

Annotation: The problem does not define how many batches of cookies Maria wants to make, making it unsolvable.

G.3 Accuracy Negative Example

Question: Ms. Carter has 378 pencils that she wants to divide equally among 14 classrooms. Each classroom should get the same number of pencils. How many pencils will be left after distributing them equally?

Solution:
To find out how many pencils will be left, we need to divide the total number of pencils (378) by the number of classrooms (14).
$378 \div 14 = 27$ with a remainder.
Let's calculate it step by step:
$14 \times 27 = 378$
But let's check how much is exactly divisible:
$14 \times 27 = 378$
Since $14 \times 27 = 378$, that means all 378 pencils are used up.
Wait — let's divide correctly:
$378 \div 14 = 27$, because $14 \times 27 = 378$.
So, 378 $(14 \times 27) = 378$ $378 = 0$.

Variable	School 1 (customization at class level)	School 2 (customization at individual level)
Economically Disadvantaged	27.4	1.4
English Language Learner	18.3	6.9
Race or Ethnicity		
White	45.8	62.8
Black	10.7	5.5
Asian	12.2	15.2
Hispanic	22.5	12.4
American Indian	0.2	0.0
Native Hawaiian	0.2	0.0
Mixed	8.4	4.1

Table 12: Demographics for students in both experimental conditions reported at the school level.

Grade Level	Number of Students	Number of Worksheets/Student
3	2	1
4	7	2
5	3	1

Table 13: Number of students and number of worksheets per student by grade level for the second experimental condition in the student study (customization at the individual level).

But let's double-check the multiplication:
$14 \times 20 = 280$
$14 \times 7 = 98$
$280 + 98 = 378$
Yes, $14 \times 27 = 378$
So, $378 \div 14 = 27$ with no remainder.
But the problem asks how many are left — so if all are distributed, the remainder is 0.
Wait — let's do the division carefully:
$378 \div 14$
14 goes into 37 two times ($14 \times 2 = 28$), subtract:
$37 \quad 28 = 9$
Bring down the 8 $\rightarrow$ 98
14 goes into 98 seven times ($14 \times 7 = 98$), subtract:
$98 \quad 98 = 0$
So, $378 \div 14 = 27$ with a remainder of 0.
Therefore, there are 0 pencils left.
But let's check: $14 \times 27 = 378$ — yes, exactly.
So, all pencils are distributed.
The final answer is 0.

Annotation: This solution is inaccurate because it conducts many unnecessary checks to ensure the final answer is correct, leading it to be inappropriate to present to a student/teacher.

G.4 Educational Appropriateness Negative Example

Question: Ms. Applebaum is organizing a field trip for her 4th-grade class. She has 25 permission slips to collect, but 7 students have already returned theirs. She wants to divide the remaining permission slips equally among the 13 students who still need to turn them in. What fraction of a permission slip does each student need to turn in?

Annotation: This question is not educationally appropriate because it is not based in reality. It is not possible for students to turn in a fraction of a permission slip.

G.5 Standards Alignment Negative Example

Grade Level: 4

Standard: Single-step or multi-step multiplication with whole numbers using a) a two-digit factor and a one-digit factor, b) a three-digit factor and a one-digit factor or c) a two-digit factor and a two-digit factor

Question: The school is preparing for the annual book fair. They ordered 24 boxes of books, and each box contains 15 books. If the school wants to arrange the books into displays with 8 books per display, how many displays can they make?

Solution:
First, we need to find the total number of books ordered. To do this, we multiply the number of boxes (24) by the number of books in each box (15).
24 x 15 = 360
So, there are 360 books in total.
Next, we need to determine how many displays they can make. To do this, we divide the total number of books (360) by the number of books per display (8).
$360 \div 8 = 45$
Therefore, the school can make 45 displays.

The final answer is 45.

Annotation: This MWP is not standards aligned because the standard calls only for multiplication and the solution requires division in addition to multiplication.

H Example Problem from ASDIV Subset

Below is an example 4th grade MWP from the labeled ASDIV 3rd-5th grade subset. The question comes directly from the ASDIV dataset, while the solution is generated by Llama 3.3 70B IT using the prompt in Appendix I.2.

Example ASDIV 3rd-5th Grade Subset MWP
Grade: 4

Standard: Single-step division with remainders for a one-digit divisor and a two- or three-digit dividend

Question: Jawbreakers come in packages of 8. Sarah ate 20 Jawbreakers. How many Jawbreakers does she have left?

Solution:
To find out how many Jawbreakers Sarah has left, we first need to determine how many packages of Jawbreakers she bought.
Since each package has 8 Jawbreakers, we need to divide the total number of Jawbreakers she ate (20) by the number of Jawbreakers in each package (8).
20 / 8 = 2 with a remainder of 4
This means she ate 2 and a half packages of Jawbreakers.
Since you cannot buy half a package, this means she must have bought at least 3 packages to eat 20 Jawbreakers (2 packages would only have 16 Jawbreakers).
To find how many Jawbreakers she has left, we need to multiply the number of Jawbreakers per package (8) by the number of packages she bought (3) and subtract the total number she ate (20)
8 x 3 = 24
24 - 20 = 4
The final answer is 4.

I Prompts

I.1 Standards Annotation Prompt for ASDIV Subset

Figure 8 displays the prompt for annotating the ASDIV 3rd-5th grade subset for standards and substandards using Llama 3.3 70B IT to ensure that each listed standard/substandard is incorporated in the problem, where the standards/substandards contain the direct text from the Virginia Standards of Learning (SOLs). We use a temperature of 0 to ensure the model selects the most probable next token during generation.

I.2 Solution Generation Prompt for ASDIV Subset

Figure 9 displays the prompt for generating step-by-step, readable solutions for the ASDIV 3rd-5th grade subset using Llama 3.3 70B IT. The few-shot examples are consistent across prompts and were manually constructed by our research team member with K-12 teaching experience. We use a temperature of 0 to ensure the model selects the most probable next token during generation.

I.3 Solution Annotation and Revision Prompts for ASDIV Subset

Figures 10 and 11 display the prompts for annotating the appropriateness of the solutions for the ASDIV 3rd-5th grade subset generated using the prompt in Appendix I.2 and revising inappropriate solutions, respectively. Both prompts are passed to Gemma 3 27B IT and the few-shot examples are consistent across prompts and were manually constructed by our research team member with K-12 teaching experience. We use a temperature of 0 for both prompts to ensure the model selects the most probable next token during generation.

I.4 MWP Generation Prompt

Figure 12 displays the standard prompt used for generating MWPs throughout this study. The 8-shot examples are filled in with randomly-selected examples for the particular grade level and standard for the query. STEM is used for the few-shot examples in the experiments in Section 4, while the ASDIV 3rd-5th grade subset is used for the few-shot examples when generating annotation data in Section 3.2. For standards that were not in the ASDIV 3rd-5th grade subset when generating annotation data, the few-shot examples were populated with 8 randomly-selected samples from the same grade

You are an experienced elementary school teacher. You are tasked with reading an educational standard(s) and its substandards and then assessing a word problem and its solution to determine whether the problem meets the standard(s) and substandards. When responding, say "Yes." or "No." to indicate whether the problem meets the specified standard(s). Please put detailed reasoning after "Yes." or "No." You only need to say "Yes." or "No." once to indicate whether the problem meets the standard(s) and substandards as a whole. Only say "Yes." if the problem exactly matches the operations and constraints mentioned in the standard(s) and substandards. For example, if the standard mentions multiplication within a certain number range, then the problem will only meet that standard if it requires multiplication within that range.

Standards: {insert standards}

Substandards: {insert substandards}

Word Problem: {insert word problem}

Solution: {insert solution}

Figure 8: Prompt for annotating ASDIV 3rd-5th grade subset for standards.

You are an experienced elementary school teacher. You are tasked with developing step-by-step solutions to math word problems for your students. The solutions should outline all the necessary steps, show complete work, and be written in a way a grade school student would understand. Make sure you separate your solution by writing “Solution:” and then the solution and end your solution with saying “The final answer is _” where “_” is filled in with the final answer. Here are some examples:

{3-shot examples}

Question: {insert question}

Solution:

Figure 9: Prompt for generating readable solutions to the ASDIV 3rd–5th grade subset.

You are an experienced elementary school teacher tasked with evaluating solutions for word problems for 3rd-5th grade students written by a less experienced teacher. The word problem and solution you will evaluate will be based on a grade level and math topic(s) and your job is to determine whether the solution is accurate, high quality, and appropriate/readable for young learners.

An accurate solution is one where the final answer and intermediate reasoning are both correct. If the final answer is correct but the intermediate reasoning is wrong, does not make sense, is unnecessarily repetitive, and/or is too complicated for a student/teacher to read, the solution is not accurate.

Here are some examples of successful evaluations:

{7-shot examples}

Now evaluate this word problem's solution and remember to answer "Yes." or "No." followed by your reasoning:

Grade Level: {insert grade level}

Math Topics: {insert math topics}

Question: {insert question}

Solution:

{insert solution}

Is the solution accurate?

Figure 10: Prompt for annotating whether generated solutions for the ASDIV 3rd–5th grade subset are appropriate.

level as the standard. For all models, we use a temperature of 1 to strike a balance between creativity and selecting the most probable next token when generating MWPs.

When customizing MWPs to student interests in Section E, the formatted string "about {topic}" is added to the prompt such that the first sentence after the 8-shot examples reads, "Write a new question about {topic} based on the grade level and math topic(s) below." In this string, "{topic}" is filled in with a random selection from the list of topics reported in Appendix I.7.

I.5 Gemma 3 27B IT Annotator Prompt and Sample Output

Figure 13 displays the prompt for annotating MWPs using Gemma 3 27B IT. The 25-shot examples are consistent across all model queries and are examples our research team manually constructed when aligning the automated annotator with expert feedback over multiple rounds of prompt iteration. We use a temperature of 0 to ensure the model selects the most probable next token during generation to improve accuracy and reproducibility of model annotations. Below is an example output for a MWP that did not MaC according to the annotator.

Example Gemma 3 27B IT Annotator Output
Grade: 3

Math Topic:
Single-step or multi-step addition and/or subtraction with whole numbers less than 1,000

Question: Maria is making bracelets for her friends. She has 235 beads. She wants to make 6 bracelets and use the same number of beads on each bracelet. How many beads will be left over after she makes the bracelets?

Solution:
Maria has 235 beads and wants to make 6 bracelets.

To find out how many beads will be left over, we need to divide the total number of beads (235) by the number of bracelets (6).

You are an experienced elementary school teacher tasked with rewriting solutions for word problems for 3rd-5th grade students written by a less experienced teacher. The word problem and original solution will be based on a grade level and math topic(s) and will contain reasoning identifying what is wrong with the original solution. Your job is to use the reasoning to rewrite the solution to ensure it is accurate, high quality, and appropriate/readable for young learners.

An accurate solution is one where the final answer and intermediate reasoning are both correct, there is not unnecessary repetition, and the wording is not too complicated for a student/teacher to read.

Here are some examples of successful solution rewrites:

{5-shot examples}

Now rewrite this word problem's solution:

Grade Level: {insert grade level}

Math Topics: {insert math topics}

Question: {insert question}

Solution:

{insert solution}

Reasoning for why the solution is inaccurate:

{insert reasoning}

Corrected Solution:

Figure 11: Prompt for revising inappropriate solutions for the ASDIV 3rd–5th grade subset as flagged using the prompt in Figure 10.

You are an experienced teacher tasked with writing word problems and solutions for 3rd-5th grade students. The question you write will be based on a grade level and math topic(s). The question's content should exactly match and incorporate ALL of the mathematical topics and constraints listed in the math topic(s). The question and answer pair you write should be solvable with the information presented in the question, contain an accurate solution, and contain language and context appropriate for a 3rd-5th grade student in a school setting (i.e., no harmful language and topics should be appropriate for school settings).

Here are some examples:

{8-shot examples}

Write a new question based on the grade level and math topic(s) below. Make sure to incorporate all of the information in the math topic(s) into your question; for example, if the math topic(s) mention remainders, the problem you write should require a remainder. Also make sure to write "Question:" followed by your question and "Solution:" followed by your solution.

Grade Level: {insert grade level}

Math Topics: {insert math topics}

Question:

Figure 12: Standard prompt for generating new MWPs.

235 ÷ 6 = 39 with a remainder of 1.

This means Maria can put 39 beads on each bracelet and she will have 1 bead left over.

The final answer is 1.

Annotation: No. While the question is solvable, has an accurate solution, and is educationally appropriate, it is not standards aligned. Specifically, the question only incorporates division and does not include addition or subtraction as stated in the math topic.

You are an experienced elementary school teacher tasked with evaluating word problems and solutions for 3rd-5th grade students written by a less experienced teacher. The word problem and solution you will evaluate will be based on a grade level and math topic(s) and your job is to determine whether the problem and solution are high quality.

There are four criteria you will use to evaluate the word problem: solvability, accuracy, educational appropriateness, and standards alignment. Questions that meet all four criteria are labeled as high quality. Any word problem or solution that does not meet one or more of the criteria is labeled as not high quality. Here is more information about how to evaluate a word problem and solution based on the four criteria:

Solvability:

A solvable question means that it can be solved with the information present and does not contain a mathematical scenario that is impossible (e.g., giving away more money or items than you have).

Accuracy:

An accurate solution is one where the final answer and intermediate reasoning are both correct. If the final answer is correct but the intermediate reasoning is wrong, does not make sense, is unnecessarily repetitive, and/or is too complicated for a student/teacher to read, the solution is not accurate.

Educational Appropriateness:

An educationally appropriate question is one you would feel comfortable giving to a student in a 3rd-5th grade school setting. Educationally appropriate questions contain content and context appropriate for students in a school setting. There are four main reasons why a question would be educationally inappropriate:

1. It contains material inappropriate for a school setting (e.g., language about harming someone)

2. It is strange, confusing, contains conflicting information, and/or is not based in reality (e.g., contains misinformation)

3. It requires no mathematical operations to solve because it gives the answer away

4. It is inappropriate for a different reason

Standards alignment:

A standards aligned question is one that adequately addresses important elements from EACH pre-specified numbered math topic. If more than one math topic is listed, then the question should incorporate important elements of EACH numbered math topic. If only one math topic is included, then it is okay if the question only incorporates elements of that topic. You should only evaluate whether the question incorporates elements from EACH listed topic; you should not penalize questions that could incorporate other topics that are not in the numbered list of topics. If a specific numbered topic lists multiple mathematical operations like addition, subtraction, and/or division, it is okay if the problem just addresses one of those operations; if a topic says "the question may include OTHER TOPIC," then

it is okay if the question does not include that other topic, as it is optional. However, the problem should incorporate meaningful elements of EACH numbered topic; for example, if a numbered topic lists decimal division, the problem should incorporate decimal division. There are four main reasons why a question would not be standards aligned:

1. It is too hard for the given topic(s)

2. It does not address some important parts of the numbered topic(s) or one or more of the numbered topic(s)

3. It does not address the numbered topic(s) at all

4. It requires additional math topics or operations that are not listed in the specified math topic(s)

Here are some examples of successful evaluations:

{25-shot examples}

Now evaluate this word problem and remember to answer "Yes." or "No." followed by your reasoning.

Grade Level: {insert grade level}

Math Topics: {insert math topics}

Question: {insert question}

Solution:

{insert solution}

Is this question high quality?

Figure 13: Prompt for annotating MWPs with Gemma 3 27B IT.

I.6 Gemma 3 27B IT Prompt for Identifying Failure Types in Annotated MWPs

Figure 14 displays the prompt for labeling the error types in model-annotated MWPs using Gemma 3 27B IT. The 8-shot examples are consistent across model queries and were manually constructed by our research team. We use a temperature of 0 to ensure the model selects the most probable next token during generation.

You are an experienced elementary school teacher tasked with reading an evaluation of a word problem and solution for 3rd-5th grade students written by a less experienced teacher and identifying the main error in the problem and/or solution that they noticed. You will have acess to the grade level, math topic, question, solution, and the less experienced teacher's evaluation of the word problem. Your job is to read the evaluation and identify which of the four criteria below it failed to incorporate as well as which specific error it made for the criteria it failed to incorporate. If the word problem violates more than 1 criteria according to the evaluation, your job is to identify the most important criteria the word problem failed to incorporate. The criteria and associated errors are:

Solvability:

A solvable question means that it can be solved with the information present and does not contain a mathematical scenario that is impossible (e.g., giving away more money or items than you have). A question is not solvable for one of two reasons:

1. It cannot be solved with the information present

2. It contains a mathematical scenario that is impossible

Accuracy:

An accurate solution is one where the final answer and intermediate reasoning are both correct. A solution is inaccurate if:

1. The final answer is wrong

2. The final answer is correct but the intermediate reasoning is wrong, does not make sense, is unnecessarily repetitive, and/or is too complicated for a student/teacher to read

Educational Appropriateness:

An educationally appropriate question is one you would feel comfortable giving to a student in a 3rd-5th grade school setting. Educationally appropriate questions contain content and context appropriate for students in a school setting. There are four main reasons why a question would be educationally inappropriate:

1. It contains material inappropriate for a school setting (e.g., language about harming someone)

2. It is strange, confusing, contains conflicting information, and/or is not based in reality (e.g., contains misinformation)

3. It requires no mathematical operations to solve because it gives the answer away

4. It is inappropriate for a different reason

Standards alignment:

A standards aligned question is one that adequately addresses important elements from EACH pre-specified numbered math topic. There are four main reasons why a question would not be standards aligned:

1. It is too hard for the given topic(s)

2. It does not address some important parts of the numbered topic(s) or one or more of the

numbered topic(s)

3. It does not address the numbered topic(s) at all

4. It requires additional math topics or operations that are not listed in the specified math topic(s)

Here are some examples of successful evaluations:

{8-shot examples}

Now evaluate this word problem and remember to both identify which criteria the problem failed to incorporate using the "Error Type:" heading and the specific error it made for that criteria using the "Specific Error:" heading. The text in the "Specific Error:" heading should exactly match one of the error types for the criteria defined above. Do not include any information other than the specific error type in the "Specific Error:" heading. Make sure to include BOTH the error type and the specific error in your response.

Grade Level: {insert grade level}

Math Topics: {insert math topics}

Question: {insert question}

Solution:

{insert solution}

Evaluation: {insert model annotation}

Error Type:

Figure 14: Prompt for labeling error types in model-annotated MWPs with Gemma 3 27B IT.

I.7 Gemma 3 27B IT Student Interest Annotation Prompt

Figure 15 displays the Gemma 3 27B IT prompt for annotating whether a generated MWP successfully incorporates a randomly-selected student interest from the list below. We use a temperature of 0 to ensure the model selects the most probable next token during generation. The list of student interests contains the 43 topics introduced in Christ et al. (2024) as well as 145 additional topics obtained by prompting GPT 4o to generate a list of topics 3rd-5th grade students would be interested in to see whether our models are able to incorporate a diverse range of student interests into generated MWPs.

Topics: ['Superman', 'Batman', 'Wonder Woman', 'Barbie', 'Power Rangers', 'basketball', 'soccer', 'football', 'volleyball', 'field hockey', 'Fortnite', 'Spiderman', 'Iron Man', 'Captain America', 'Captain Marvel', 'Thor, the God of Thunder', 'Black Panther', 'Taylor Swift', 'swimming', 'Pokémon', 'Super Mario', 'Naruto', 'unicorns', 'Hello Kitty', 'Minecraft', 'lacrosse', 'cheer leading', 'LeBron James', 'Steph Curry', 'Patrick Mahomes', 'Serena Williams', 'dogs', 'cats', 'dinosaurs', 'Harry Potter', 'cars', 'planes', 'trains', 'pizza', 'cookies', 'ice cream', 'candy', 'Frozen (Elsa and Anna)', 'Star Wars', 'Paw Patrol', 'My Little Pony', 'Minions', 'Jurassic Park', 'SpongeBob SquarePants', 'Disney Princesses', 'Toy Story', 'The Incredibles', 'Scooby-Doo', 'Peppa Pig', 'Dora the Explorer', 'Pikachu', 'Thomas the Tank Engine', 'Sonic the Hedgehog', 'Transformers', 'Cinderella', 'Moana', 'Shrek', 'Winnie the Pooh', 'Tom and Jerry', 'Sesame Street', 'The Lion King', 'Alice in Wonderland', 'The Little Mermaid', 'Peter Pan', 'Aladdin', 'The Jungle Book', 'Pocahontas', 'Beauty and the Beast', 'Frozen', 'Ratatouille', 'Finding Nemo', 'Cars', 'Up', 'The Simpsons', 'Looney Tunes', 'Teenage Mutant Ninja Turtles', 'Mythical Creatures (dragons, unicorns)', 'Dinosaurs', 'Space and Astronauts', 'Robots', 'Aliens', 'Exploring the Ocean', 'Underwater Creatures', 'Pirates', 'Fairies', 'Wizards', 'Magic Tricks', 'Time Travel', 'Detectives and Mystery', 'Inventions', 'The Avengers', 'The Justice League', 'Dance and Ballet', 'Music Instruments', 'Art and Drawing', 'Science Experiments', 'Cooking and Baking', 'DIY Crafts', 'Board Games', 'Puzzles', 'Riddles', 'Pets (cats, dogs, hamsters)', 'Farm Animals', 'Zoo Animals', 'Wildlife Conservation', 'Plants and Gardening', 'Hiking and Nature', 'Weather and Meteorology', 'The Solar System', 'Camping', 'National Parks', 'Trains and Railroads', 'Planes and Aviation', 'Cars and Racing', 'Construction Vehicles', 'Firefighters', 'Police Officers', 'Doctors and Nurses', 'Astronauts and Space Exploration', 'Animals and Wildlife', 'Space and Astronomy', 'Robots and Technology', 'Underwater Life', 'Fairy Tales and Folklore', 'Outer Space', 'Music and Instruments', 'Insects and Bugs', 'Historical Figures', 'Countries and Cultures', 'Mythical Creatures', 'Magic and Wizards', 'Friendship and Relationships', 'Ocean Life', 'Cars and Vehicles', 'Famous Inventors', 'Famous Artists', 'Ancient Civilizations', 'Space Exploration', 'Gardening', 'Environmental Conservation', 'Pirates and Treasure', 'Famous Scientists', 'Computer Programming', 'Unexplained Mysteries', 'Planets and the Solar System', 'Cartoons and Animated Shows', 'Photography', 'Books and Reading', 'Volcanoes', 'Mythology', 'Ancient Egypt', 'Reptiles and Amphibians', 'Recycling', 'Fairy Gardens', 'Indoor Games', 'Marine Biology', 'Virtual Reality', 'Natural Disasters', 'Construction and Building', 'the Circus and Performing Arts', 'Science Fiction', 'Pottery and Ceramics', 'Famous Explorers', 'Birds and Bird Watching', 'Famous Landmarks', 'Health and Nutrition', 'Myths and Legends', 'Fashion and Clothing', 'DIY Science Projects', 'Cultural Festivals', 'Forests and Trees', 'Mummies', 'Famous Composers', 'Circus Animals', 'Geology', 'Farm Life', 'Travel and Adventure', 'Ballet and Dance', 'Whales and Dolphins', 'Mystery Stories', 'Hiking and Camping', 'Games and Puzzles', 'Space Aliens and UFOs']

Your goal is to read a math word problem and determine whether it effectively includes a pre-specified topic. If the question does not effectively include the topic, write "No." followed by your reasoning. If the question does effectively include the topic, write "Yes." followed by your reasoning.

Now evaluate whether this problem includes the specified topic and remember to exactly answer "Yes." or "No." followed by your reasoning.

Topic: {insert topic}

Question: {insert question}

Does the question effectively incorporate the specified topic?

Figure 15: Prompt for annotating whether MWPs successfully include a pre-specified topic.