Math Word Problem Generation and Evaluation Using Large Language Models

Anonymous submission

Abstract

Creating exercises and quizzes is a time-consuming and repetitive task for teachers aiming to assess students' mastery of knowledge components (KCs). Large language models (LLMs), which have demonstrated strong capabilities in mathematical reasoning, hold great promise for automating the generation of math word problems (MWPs), thereby allowing teachers to focus more on providing support to students in need. While LLMs' problem-solving abilities on MWPs have been widely studied, their capacity to generate high-quality MWPs has received relatively less attention. In this paper, we investigate whether two state-of-the-art LLMs—GPT-4.1 and Claude Sonnet 4—can generate pedagogically sound, curriculum-aligned, and cognitively appropriate MWPs suitable for primary and secondary school students. To increase the challenge, we task the models with generating problems that integrate two KCs, requiring meaningful synergy between them. The generated MWPs are evaluated through both human and LLM assessments based on a set of task-oriented and linguistic criteria. Our experimental results show that, as MWP generators, LLMs can produce problems that meet individual evaluation criteria in most cases (exceeding 80% for all but one criterion). As evaluators, GPT-4.1 tends to overrate the quality of generated problems, while Claude Sonnet 4 is more likely to wrongly flag problems as failing to meet criteria. These findings highlight the importance of developing rigorous AI evaluators, as both a capable generator and a reliable evaluator are essential to fully automate MWP generation.

Introduction

The rapid advancement of large language models (LLMs) in recent years has transformed various sectors, including education. LLMs are widely regarded as having great potential to revolutionize teaching and learning by automating the generation of educational content and enabling more interactive, conversational, and adaptive learning experiences (Giannakos et al. 2025). In particular, if LLMs can generate high-quality content aligned with curriculum standards, student proficiency levels, and specific learning objectives, they could significantly reduce educators' workload, allowing teachers to dedicate more time to understanding and supporting students who need additional help.

Large language models (LLMs) have demonstrated strong capabilities in mathematical reasoning and offer significant potential for automating the generation of math word problems (MWPs). Although their problem-solving abilities on MWPs have been extensively studied (Wang et al. 2025; Forootani 2025), less attention has been paid to their ability to generate high-quality MWPs.

In this study, we explore the use of two state-of-the-art LLMs—GPT-4.1 and Claude Sonnet 4—for MWP generation. These models were selected for their proven strengths in language generation, logical reasoning, and strict adherence to prompts, qualities that are essential for producing high-quality educational content. Furthermore, both models place a strong emphasis on safety and alignment in their outputs, making them particularly suitable for use in educational settings. We task the models with generating MWPs based on math knowledge components (KCs) drawn from primary and secondary school curricula. To increase the challenge, we prompt the models to create MWPs that integrate a pair of KCs within a given topic, requiring meaningful synergy between the two KCs within the topic context.

In addition to examining the performance of the two models in generating MWPs, we also evaluate their ability in assessing the quality of MWPs. Specifically, both LLMs are used to evaluate MWPs based on six task-oriented criteria and four linguistic criteria. Ground-truth evaluations are established through two rounds of manual assessment. The first round includes cross-rater calibration, alignment, and independent evaluations by multiple raters. In the second round, a senior team member collaborates with the initial raters to review and refine the assessments, taking into account the evaluations provided by the LLMs. We then assess the performance of the two models as evaluators by measuring their agreement with the ground-truth assessments.

The main contributions and findings of our paper are summarized as follows:

- We investigate the performance of two state-of-the-art LLMs in generating MWPs for primary and secondary school levels, focusing on problems that integrate pairs of math KCs on a given topic. Both LLMs are able to satisfy most evaluation criteria in the majority of cases. The most challenging criterion is *Synergy*, which requires meaningful integration of the two KCs: only 69% of MWPs generated by the two models meet this criterion. We also examine the diversity of generated MWPs. While MWPs across different topics are well differenti-

ated, diversity within the same topic remains limited and warrants further improvement.

- We propose a comprehensive evaluation framework consisting of six task-oriented criteria and four linguistic criteria to assess MWP quality. Using this framework, we analyze the effectiveness of the two LLMs as evaluators. Both models exhibit lower accuracy and specificity compared to human raters. Claude Sonnet 4 is more sensitive to low-quality MWPs, often flagging issues that human raters overlook, but this comes at the cost of reduced specificity. GPT-4.1, on the other hand, tends to overrate MWPs, regardless of whether the MWPs are generated by itself or not.
- We construct a dataset comprising 1,056 MWPs, each annotated with ground-truth evaluations based on the ten proposed criteria. This dataset will be made publicly available to support future research. It can serve as training or benchmark data for developing more accurate AI evaluators to enable fully automated MWP generation.

In the remainder of the paper, we first review related work, followed by a detailed description of our MWP generation and evaluation methodology, and finally present and discuss our findings.

Related Work

Math Word Problem Generation

Early approaches to MWP generation before the advent of LLMs primarily relied on template-based methods (Polozov et al. 2015; Koncel-Kedziorski et al. 2016). The MWPs generated this way follow fixed patterns and manual effort is required to construct the templates. Neural models are trained in (Zhou and Huang 2019; Wang, Lan, and Baraniuk 2021) to generate MWPs from given equations and topics without predefined templates, but they focus on single equations.

Recent research has increasingly leveraged the generative capabilities of LLMs to produce diverse and pedagogically relevant MWPs. Niyarepola et al. (2022) investigated the use of multilingual pre-trained models for MWP generation. MATHWELL (Christ, Kropko, and Hartvigsen 2024) fine-tune models on teacher-annotated data to generate K-8 level MWPs without a reference MWP or equation. Ariyarathne et al. (2025) generated elementary MWPs using the number of MWPs needed, the grade and type of question (e.g. addition, subtraction) as inputs, and found that LLMs struggle to generate MWPs that adhere to the given grade and question type. Compared with these work, the KCs used in our work are much more fine-grained and better aligned with standard math curricula.

LLMs have also been used to generate diverse and more challenging MWPs to improve the mathematical reasoning capabilities of LLMs such as in (Cao et al. 2021; Zhou et al. 2023; Kang et al. 2024; Chen et al. 2024). The requirements on these MWPs are different from those generated for education purpose.

Evaluation of Math Word Problems

Earlier evaluations of MWPs often largely relied on automatic metrics such as BLEU (Papineni et al. 2002) or ROUGE (Lin 2004). Domain-specific metrics such as equation relevance, topic relevance, solvability have been introduced in (Zhou and Huang 2019; Wang, Lan, and Baraniuk 2021; Christ, Kropko, and Hartvigsen 2024) to better reflect pedagogical alignment, where equation relevance is similar to KC alignment in our work.

Scaria, Chenna, and Subramani (2024) investigated the ability of five LLMs to generate questions of different cognitive levels, as defined by Bloom's taxonomy, for a data science course. A nine-item rubric is used to assess the linguistic and pedagogical quality of the generated questions including *Understandable*, *TopicRelated*, *Grammatical*, *Clear*, *Answerable* and *Bloom's Level etc.*. Fu et al. (2024) introduced QGEVAL, a framework for evaluation questions generated on a given paragraph and answer across 7 dimensions: fluency, clarity, conciseness, relevance, consistency, answerability, and answer consistency. Our work builds on these two evaluation framework and extends them for evaluating MWPs.

To the best of our knowledge, this is the first study to investigate the capabilities of LLMs in generating MWPs based on a given pair of KCs and a topic for educational purposes, as well as their ability to assess MWPs using a comprehensive evaluation framework comprising six task-oriented and four linguistic criteria

Math Word Problem Generation

We generate math word problems (MWPs) using math knowledge components (KCs) from both primary and secondary school curricula. Our initial experiments show that GPT-4.1 and Claude Sonnet 4 perform well when generating MWPs based on a single KC. To increase the complexity, we extend the task by requiring the models to generate MWPs that integrate a pair of KCs along with a given topic. These KC pairs must be combined in a meaningful and coherent way. In this section, we describe the selected KC pairs, the associated topics, and the specific requirements for the MWPs.

Knowledge component pairs

The math knowledge components (KCs) used in our study are selected from the mathematics syllabi published by the Ministry of Education (MOE). Certain KCs, such as those related to geometry, require visual components to form complete problems and are therefore excluded from this study, as we focus specifically on word problems. We made minor edits to the original descriptions of the KCs to improve clarity and ensure that the scope and intent of each KC are more easily understood by LLMs. In total, we curated and selected 117 KCs for this study. We organize the selected KCs using a three-level hierarchy:

- **Level 0**: Number types, including whole numbers, fractions, decimals, percentages, rates, ratios, algebra, statistics and probability, bases and powers, sets, and matrices.
- **Level 1**: Operations, including representation and concept, addition, subtraction, multiplication, division, finding, comparison and ordering, conversion, rounding, evaluation, simplification, and solving.

- **Level 2**: A detailed description of each KC, including the complexity of the numbers involved and further elaboration on the associated operations.

A KC pair consists of a **primary** KC and a **secondary** KC. The primary KC determines the grade level and overall difficulty of the generated math word problems. The secondary KC is selected from the same grade or a lower grade and adds an extra layer of complexity to the problem. This pairing strategy introduces greater structural complexity and allows for a more rigorous evaluation of the generative capabilities of large language models (LLMs). Our goal is to assess whether LLMs can effectively integrate two distinct yet conceptually related KCs in a coherent manner—such that one KC builds upon or contributes to the solution path of the other—an aspect that, to the best of our knowledge, has not been vigorously examined in prior work.

We observed that certain KCs are intrinsically incompatible with one another. Pairing such KCs often led to incoherent or awkward word problems—not due to limitations of the LLMs, but rather due to fundamental conceptual mismatches between the KCs themselves. To address this issue, we established a set of pairing rules to guide the selection of compatible KC combinations. For example, KCs involving operations such as comparison and ordering, conversion, rounding, evaluation, and simplification were restricted to being used only as secondary KCs, as they typically occur in the final steps of problem solving. Additionally, these secondary KCs could only be paired with primary KCs that involve the same number types. After applying these rules, we conducted a manual review and selected 176 coherent and compatible KC pairs for inclusion in our study. Each of the 117 KCs appears in at least one of these pairs. The pairing rules can be found in supplementary materials.

Topics

We instructed the two LLMs to generate MWPs based on three real-world topics: household finance, recreation, and services. These topics were selected due to their strong relevance to everyday life, which helps ensure the contextual familiarity and practical applicability of the generated MWPs.

- **Household finance** such as income, utility bills, money, interest, savings, installment, mortgage, financial planning etc.
- **Recreation** such as sports, games, exercises, music, movies, dancing, painting, fishing and other recreational activities.
- **Services** such as installation, maintenance, repairing, cleaning, laundry, hotel, retail, e-commerce, streaming services, digital services etc.

Requirements on generated MWPs

We evaluate the quality of the generated MWPs using both task-oriented and linguistic criteria. The task-oriented criteria include the following:

- **Answerability:** The word problem must be answerable with no missing information, no conflicting information and no illogical relationship.
- **Primary KC Alignment:** The word problem must be on the intended primary KC, that is, the primary KC is needed for solving the word problem. The primary KC decides the main idea and difficulty level of the word problem and there should be no other KC in the word problem that is harder than it.
- **Secondary KC Alignment:** The word problem must also target the intended secondary KC, that is, the secondary KC is also needed for solving the word problem. The secondary KC adds complexity to the word problem.
- **Synergy:** Primary KC and secondary KC need to be used in tandem to solve the word problem. If there are multiple questions in the word problem, the questions need to be connected with each other such that the result of an early question should be used as known information to a later question in the word problem.
- **Topic Alignment:** The context of the word problem must belong to the given topic.
- **Real-World Plausibility:** The word problem must describe a realistic scenario mirroring how people typically act and operate in reality.

Linguistic criteria include the following four dimensions:

- **Clarity:** The language used in the word problem must be easy to understand and free from ambiguity.
- **Conciseness and Relevance:** There is no irrelevant information in the word problem that is not useful for solving the problem. All given numbers must be necessary for solving the word problem.
- **Language Quality:** The word problem uses correct grammar, spelling, and vocabulary.
- **Content Appropriateness:** The word problem is respectful, age-appropriate, and free from offensive content.

The generated MWPs are required to satisfy all the 10 criteria. To ensure this, all requirements are explicitly included in the prompts provided to the two LLMs. The prompts used for MWP generation consist of four sections:

- **Role and Objective:** Defines the role of the LLM as an educational content generator for mathematics and the task as generating MWPs based on a given KC pair and topic.
- **Requirements:** Specifies the six task-oriented criteria and four linguistic criteria that each generated MWP must fulfill, along with a requirement that the MWP needs to be formatted in LaTeX.
- **Input:** Provides the primary KC, secondary KC, topic, and grade level as input parameters.
- **Output Format:** Instructs the model to return a JSON object with two fields: the MWP and its corresponding solution.

The first three sections of the prompts are kept largely identical for both LLMs to ensure a fair comparison. The output format section is adapted to better align with the respective output conventions of each model. Full prompts to the two models are provided in the supplementary materials.

For each combination of KC pair and topic, one MWP is generated by each LLM. In total, we generate 1056 MWPs (176 KC pairs $\times$ 3 topics $\times$ 2 LLMs).

Math Word Problem Evaluation

The generated MWPs were evaluated using both LLMs and manual assessment. Each MWP was assessed according to six task-oriented criteria and four linguistic criteria, as described previously. All criteria were binary, with a score of 1 indicating a positive rating and 0 indicating a negative rating. If a problem failed the *Answerability* criterion—due to missing or conflicting information, logical inconsistencies, or unsolvability—subsequent criteria were not evaluated. This approach reflects the observation that unanswerable problems generally exhibit cascading deficiencies across multiple quality dimensions. Furthermore, if either the *Primary KC Alignment* or *Secondary KC Alignment* criterion was not satisfied, the *Synergy* criterion was likewise not assessed. This is because *Synergy* inherently relies on the correct alignment of both KCs, and without this foundational alignment, evaluating their interdependence is invalid.

Automated evaluation using LLMs

In this setup, each generated MWP is evaluated both by the model that produced it (self-evaluation) and by the other model (cross-model evaluation). The prompts provided to the two models for MWP evaluation consist of four sections:

- **Role and Objective:** defines the LLM's role as an MWP evaluator and the task as assessing quality of MWPs generated for a given KC pair and topic.
- **Evaluation Criteria:** details the six task-oriented and four linguistic criteria that the generated MWPs are required to satisfy.
- **Input:** presents the MWP along with the primary KC, secondary KC, topic, and grade level used in generation.
- **Output Format:** requires the response to be structured as a JSON object containing 11 fields—one for each of the 10 evaluation criteria, and an additional field for explaining the ratings.

To enhance robustness and reduce response variability, each MWP was evaluated five times. The final label assigned to each criterion was determined by a majority vote across these five evaluation runs.

Manual evaluation and cross-rater alignment

The manual evaluation was conducted by a panel of three raters, all of whom were university students majoring in computer science or electrical and electronic engineering. The evaluation process was carried out in multiple stages to enhance inter-rater consistency and alignment.

Stage 1: Cross-rater calibration. The three raters were provided with the 10 evaluation criteria and independently assessed the same set of 20 MWPs. Any discrepancies in their evaluations were discussed and resolved prior to proceeding to the next batch of 20 MWPs. In total, five rounds of calibration were performed, each involving 20 MWPs. The MWPs used in this stage were excluded from the final set of 1,056 MWPs.

Stage 2: Cross-rater alignment. To quantify inter-rater agreement, all three raters independently evaluated a common subset of 100 MWPs generated by GPT-4.1 on the topic of "Recreation". The percentage of agreement on all 10 criteria exceeded 90%.

Stage 3: Independent evaluation. The remaining MWPs were partitioned among the three raters, with each MWP evaluated independently by a single rater. MWPs sharing the same KC pair but differing in topic were assigned to the same rater to maintain consistency.

	ground-truths		Manual		GPT		Claude	
LLMs	high	low	high	low	high	low	high	low
GPT	300	228	363	165	425	103	301	227
Claude	286	242	379	149	422	106	344	184

Table 1: Number of high-quality (high) and low-quality (low) MWPs in ground-truths and in evaluations given by different evaluators.

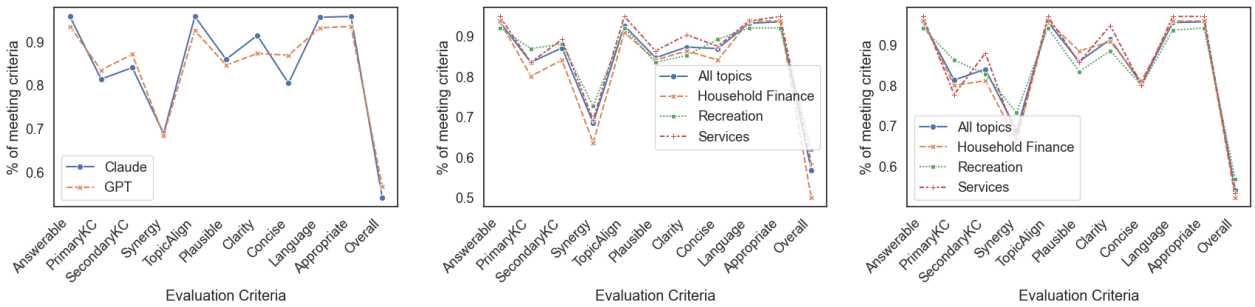
Groud-truth evaluations

After all MWPs were evaluated by both manual evaluation and LLM evaluators, MWPs that received a rating of 1 on all 10 criteria from both sources were designated as *high-quality* MWPs. All remaining MWPs—those that received at least one rating of 0 from either a human rater or an LLM—were subjected to further manual validation by a senior researcher in computer science. Any disagreements between the senior member and the initial human raters were discussed and resolved collaboratively. The resulting consensus ratings were treated as the ground-truth evaluations. If an MWP received a final rating of 1 on all criteria, it was classified as a ground-truth *high-quality* MWP; otherwise, it was considered a ground-truth *low-quality* MWP.

Table 1 presents the number of *high-quality* and *low-quality* MWPs according to the ground-truth evaluations (2nd and 3rd columns), initial manual evaluation (4th and 5th columns), GPT-4.1 (6th and 7th columns), and Claude Sonnet 4 (last two columns). The number of ground-truth *high-quality* MWPs generated by GPT-4.1 is slightly higher than that generated by Claude Sonnet 4. Both human raters and GPT-4.1 tend to overestimate MWP quality. In contrast, Claude Sonnet 4's ratings on MWPs generated by GPT-4.1 are more aligned in quantity with the ground-truth evaluations, though the overlap in specific MWPs is not high. Notably, Claude Sonnet 4 appears to favor the MWPs produced by itself.

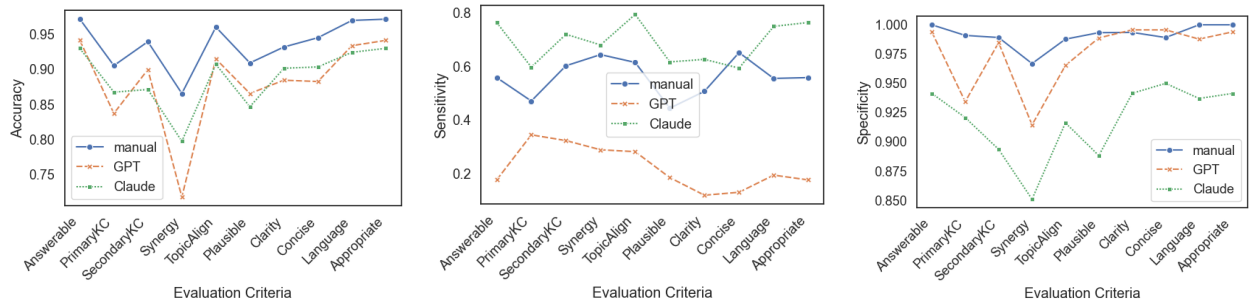
Evaluating alignment with ground-truths

To evaluate the effectiveness of LLMs as MWP evaluators, we compare their assessments against the ground-truth evaluations. In the context of math word problem generation, it is particularly important to identify MWPs that fail to meet evaluation criteria, as exposing students to such flawed problems may lead to confusion and frustration.



(a) all topics (b) GPT (c) Claude

Figure 1: Percentage of MWPs meeting individual criteria



(a) Accuracy (b) Sensitivity (c) Specificity

Figure 2: Performance of LLM and human evaluators on MWPs generated by GPT-4.1

The performance of LLM-based evaluators is assessed using three metrics: *accuracy*, *sensitivity*, and *specificity*, with a focus on detecting MWPs that do not satisfy a given criterion. Let c denote an evaluation criterion, R be the set of evaluations produced by a given evaluator, and N be the total number of MWPs. Accuracy, sensitivity, and specificity with respect to criterion c are defined as follows:

$$accuracy_c(R) = \frac{TP_c(R) + TN_c(R)}{N} \quad (1)$$

$$sensitivity_c(R) = \frac{TP_c(R)}{TP_c(R) + FN_c(R)} \quad (2)$$

$$specificity_c(R) = \frac{TN_c(R)}{TN_c(R) + FP_c(R)} \quad (3)$$

where $TP_c(R)$ is #true positives (MWPs that do not meet criterion c and are correctly identified by the evaluator), $TN_c(R)$ is #true negatives (MWPs that meet criterion c and are correctly identified by the evaluator), $FP_c(R)$ is #false positives and $FN_c(R)$ is #false negatives in R.

Accuracy measures the overall proportion of correct evaluations. *Sensitivity* (also known as recall) reflects the evaluator's ability to correctly identify MWPs that fail to meet criterion c. *Specificity* captures the evaluator's ability to correctly identify MWPs that satisfy criterion c.

Experimental Results

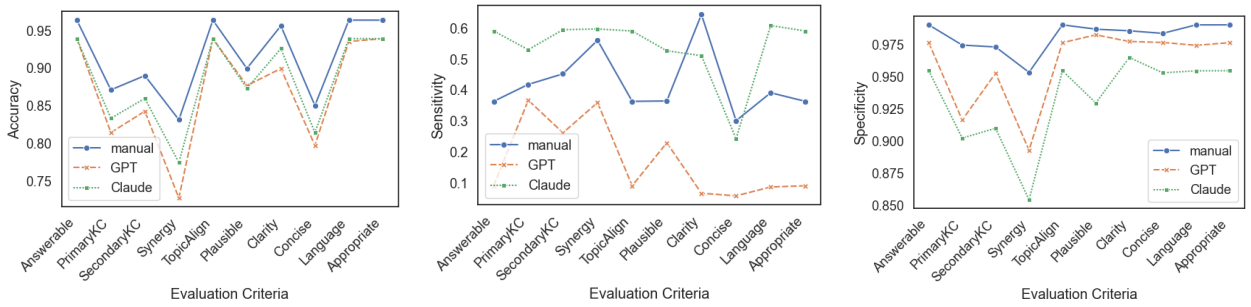
In this section, we present our findings to answer the following research questions:

- **RQ1:** What is the quality of MWPs generated by state-of-the-art LLMs based on KC pairs drawn from primary and secondary mathematics curricula?
- **RQ2:** How well do state-of-the-art LLMs perform in evaluating the quality of MWPs?
- **RQ3:** To what extent are the MWPs generated by LLMs diverse?

Quality of MWPs generated by the two models

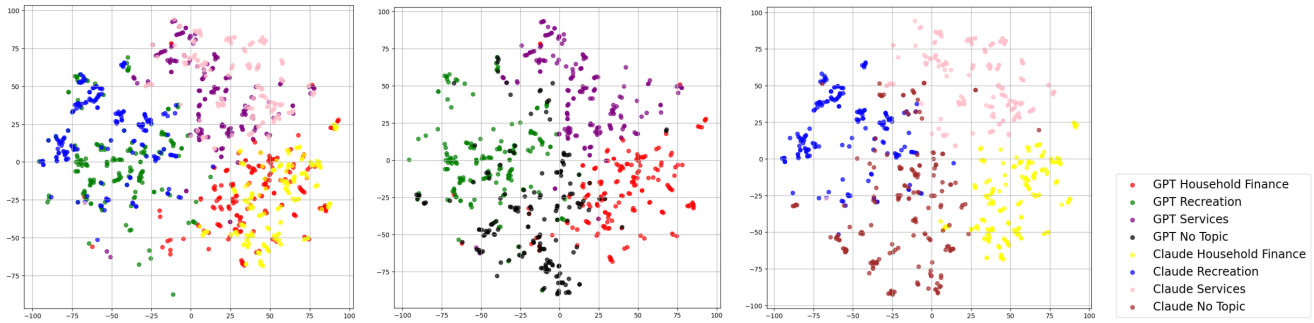
Figure 1 presents the percentage of MWPs that satisfy each individual evaluation criterion, based on ground-truth assessments. An MWP is considered to meet the *Overall* criterion—i.e., to be classified as *high-quality*—if it receives a rating of 1 on all individual criteria.

Both LLMs perform strongly on the four basic requirements—*Answerability*, *Language Quality*, *Content Appropriateness*, and *Topic Alignment*—achieving scores above 93%. They also perform well on more nuanced dimensions such as *Primary KC Alignment*, *Secondary KC Alignment*, *Real-World Plausibility*, *Clarity*, and *Conciseness and Relevance*, each exceeding 80%. However, both models exhibit notable weakness in the *Synergy* dimension, with only 69% of generated MWPs meeting this criterion. The most frequent reason for failing the *Synergy* criterion is



(a) Accuracy (b) Sensitivity (c) Specificity

Figure 3: Performance of LLM and human evaluators on MWPs generated by Claude Sonnet 4



(a) two LLMs on three topics (b) GPT (c) Claude

Figure 4: Visualization of the MWPs generated by the two LLMs on the three topics

that the sub-questions within an MWP are not meaningfully interconnected.

Among the MWPs generated by GPT-4.1, only 57% satisfy all evaluation criteria and are therefore classified as *high-quality* MWPs. For Claude Sonnet 4, the proportion of *high-quality* MWPs is slightly lower, at 54%. As shown in Figure 1(b) and Figure 1(c), the quality of the generated MWPs does not vary significantly across the three topics.

Performance of LLMs as MWP evaluators

Figure 2 and Figure 3 show the performance of the LLM evaluators and the initial manual evaluations (denoted as "manual" in the figures) on MWPs generated by GPT-4.1 and Claude Sonnet 4, respectively. Among the three evaluators, manual evaluation demonstrates the highest accuracy and specificity. High specificity indicates that manual raters are less likely to incorrectly flag an MWP that satisfies a given criterion as failing it, whereas both LLM evaluators are more prone to such false positives. Manual evaluation is limited in sensitivity, likely due to the cognitive effort required to detect more subtle or deeply embedded issues.

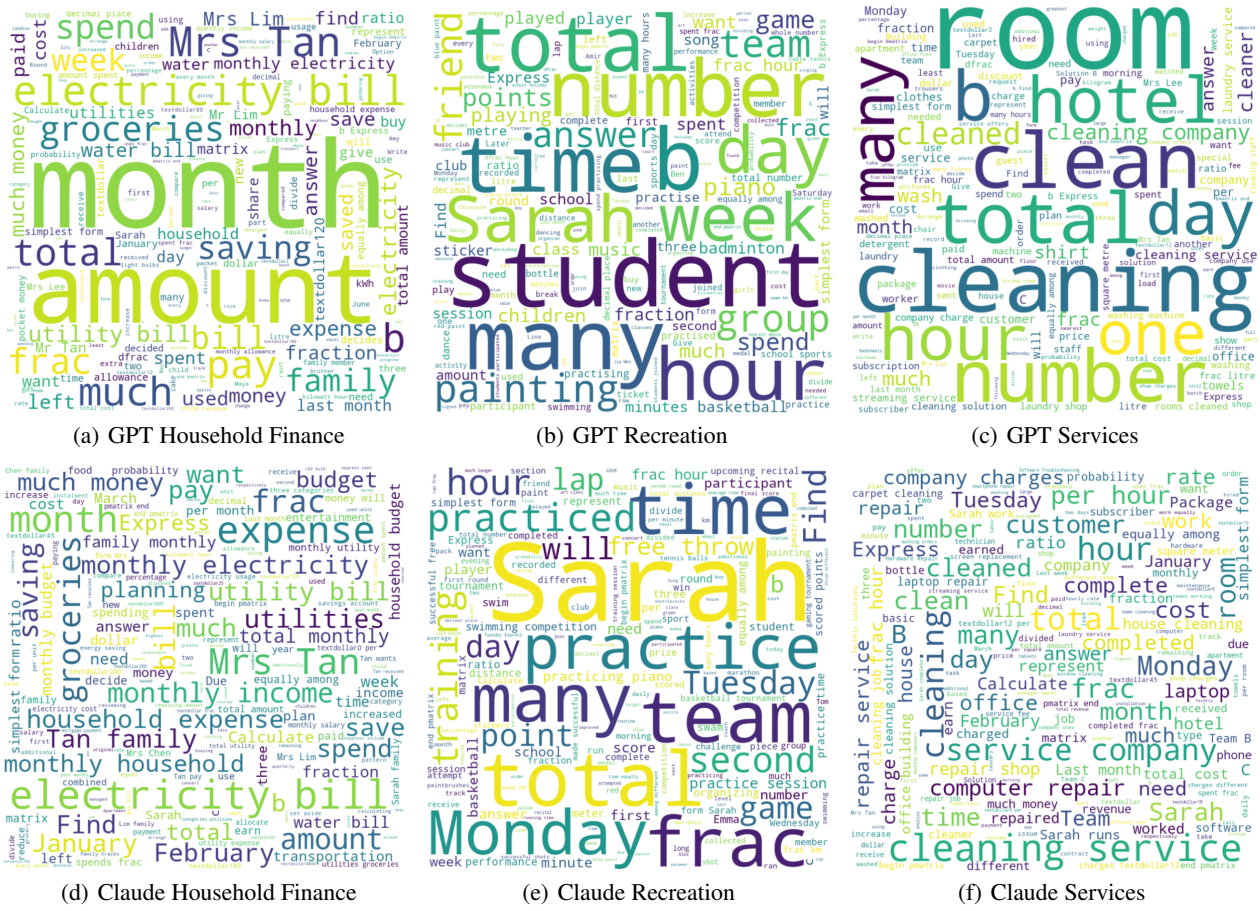
Claude Sonnet 4 achieves the highest sensitivity, even surpassing manual evaluation, though this comes at the expense of reduced specificity. In contrast, GPT-4.1 exhibits the lowest sensitivity among the three evaluators, with values generally below 0.3, indicating its limited effectiveness in identifying low-quality MWPs. Both LLMs also show notable difficulty in reliably evaluating the *Synergy* criterion.

Diversity of the generated MWPs

In this experiment, we assess the diversity of generated MWPs across different topics. We use the BGE embedding model (bge-small-en-v1.5)(Xiao et al. 2023) to encode each MWP into a high-dimensional vector representation. These vectors are then projected into a two-dimensional space using the t-SNE algorithm(Maaten and Hinton 2008) for visualization. Figure 4 shows that MWPs generated on the three topics form clearly separated clusters, indicating topic-level diversity. Furthermore, MWPs generated on the same topic by different LLMs exhibit substantial overlap in the projected space, suggesting that topic plays a more significant role in shaping the content than the choice of LLM.

Additionally, we prompted both models to generate MWPs for all KC pairs without specifying a topic. As shown in Figure 4(b) and Figure 4(c), the MWPs generated without a topic tend to fall between the Recreation and Household Finance clusters for both models. This is an interesting observation. A closer examination reveals that many of these MWPs are set in school contexts, which may be influenced by the *Role and Objective* section of the generation prompt provided to the models.

However, as shown in Figure 5, both GPT-4.1 and Claude Sonnet 4 exhibit relatively low diversity within individual topics. This limitation is particularly evident in Figure 5(b),



(a) GPT Household Finance (b) GPT Recreation (c) GPT Services

(d) Claude Household Finance (e) Claude Recreation (f) Claude Services

Figure 5: Word cloud of MWPs generated by the two LLMs on the three topics

Figure 5(c), and Figure 5(e). For the *Recreation* topic, many MWPs feature "Sarah" as the main character by both models. In the *Services* topic by GPT-4.1, most problems focus on cleaning or laundry services, with minimal coverage of other relevant subtopics such as streaming or digital services. These patterns suggest that, although LLMs are capable of aligning their outputs with the specified topic, they tend to exhibit limited creativity and variability when granted freedom over the choice of subtopics.

Summary and Conclusion

In this paper, we investigated the capabilities of two state-of-the-art LLMs—GPT-4.1 and Claude Sonnet 4—in the generation and evaluation of MWPs. Both models are able to produce MWPs that meet individual evaluation criteria with high percentage (exceeding 80% on nine dimensions). However, they struggle to generate MWPs that satisfy all criteria simultaneously, resulting in a relatively low proportion of overall high-quality problems. Regarding diversity, both models effectively adhere to topic-level instructions but show limited variability at the subtopic level when explicit guidance is absent, often defaulting to a narrow set of recurring themes.

In terms of evaluation behavior, GPT-4.1 tends to be overly optimistic, exhibiting high specificity but low sensitivity, while Claude Sonnet 4 demonstrates the opposite trend, with higher sensitivity but lower specificity. These contrasting tendencies suggest that, although LLMs show promise as supportive tools in MWP generation and assessment, they are not yet sufficiently reliable to function as fully autonomous generators or evaluators. Manual evaluation achieves the highest overall accuracy but is limited in sensitivity, likely due to the cognitive effort required to detect more subtle or deeply embedded issues.

All KCs, KC pairs, generated MWPs, and their corresponding evaluations from different evaluators will be made publicly available upon acceptance of this work. Future research can leverage this ground-truth-labeled dataset to refine prompting strategies—such as through few-shot learning—and to incorporate exemplar-based approaches aimed at enhancing the robustness of LLM-based evaluations. Additionally, the dataset can support the development of customized models for more reliable and accurate MWP evaluation. Extending the analysis to additional LLMs, particularly open-source models, may further uncover the strengths and limitations of current LLMs in MWP generation and evaluation, thereby advancing the development of robust automated problem-generation systems.

References

Ariyarathne, N.; Bandara, H.; Heshan, Y.; Gamage, O.; Ranathunga, S.; Nayanajith, D.; Sivapalan, Y.; Lihinikaduarachchi, G.; Vihidun, T.; Chandirakumar, M.; Premakumar, S.; and Gathsara, S. 2025. Elementary Math Word Problem Generation using Large Language Models. arXiv:2506.05950.

Cao, T.; Zeng, S.; Zhao, S.; Mansur, M.; and Chang, B. 2021. Generating Math Word Problems from Equations with Topic Controlling and Commonsense Enforcement. arXiv:2012.07379.

Chen, N.; Wu, N.; Chang, J.; Shou, L.; and Li, J. 2024. ControlMath: Controllable Data Generation Promotes Math Generalist Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12201–12217. Miami, Florida, USA: Association for Computational Linguistics.

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

Forootani, A. 2025. A Survey on Mathematical Reasoning and Optimization with Large Language Models. arXiv:2503.17726.

Fu, W.; Wei, B.; Hu, J.; Cai, Z.; and Liu, J. 2024. QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11783–11803. Miami, Florida, USA: Association for Computational Linguistics.

Giannakos, M.; Azevedo, R.; Brusilovsky, P.; Cukurova, M.; Dimitriadis, Y. A.; Leo, D. H.; Järvelä, S.; Mavrikis, M.; and Rienties, B. 2025. The promise and challenges of generative AI in education. *Behav. Inf. Technol.*, 44(11): 2518–2544.

Kang, X.; Wang, Z.; Jin, X.; Wang, W.; Huang, K.; and Wang, Q. 2024. Template-Driven LLM-Paraphrased Framework for Tabular Math Word Problem Generation. arXiv:2412.15594.

Koncel-Kedziorski, R.; Konstas, I.; Zettlemoyer, L.; and Hajishirzi, H. 2016. A Theme-Rewriting Approach for Generating Algebra Word Problems. In *EMNLP*, 1617–1628. The Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Niyarepola, K.; Athapaththu, D.; Ekanayake, S.; and Ranathunga, S. 2022. Math Word Problem Generation with Multilingual Language Models. In *Proceedings of the 15th International Conference on Natural Language Generation*, 144–155. Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Polozov, O.; O'Rourke, E.; Smith, A. M.; Zettlemoyer, L.; Gulwani, S.; and Popovic, Z. 2015. Personalized Mathematical Word Problem Generation. In *IJCAI*, 381–388. AAAI Press.

Scaria, N.; Chenna, S. D.; and Subramani, D. N. 2024. Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation. In *AIED*, volume 14830 of *Lecture Notes in Computer Science*, 165–179. Springer.

Wang, P.-Y.; Liu, T.-S.; Wang, C.; Wang, Y.-D.; Yan, S.; Jia, C.-X.; Liu, X.-H.; Chen, X.-W.; Xu, J.-C.; Li, Z.; and Yu, Y. 2025. A Survey on Large Language Models for Mathematical Reasoning. arXiv:2506.08446.

Wang, Z.; Lan, A.; and Baraniuk, R. 2021. Math Word Problem Generation with Mathematical Consistency and Problem Context Constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5986–5999. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Xiao, S.; Liu, Z.; Zhang, P.; and Muennighoff, N. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597.

Zhou, Q.; and Huang, D. 2019. Towards Generating Math Word Problems from Equations and Topics. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, 494–503. Association for Computational Linguistics.

Zhou, Z.; Ning, M.; Wang, Q.; Yao, J.; Wang, W.; Huang, X.; and Huang, K. 2023. Learning by Analogy: Diverse Questions Generation in Math Word Problem. In *Findings of the Association for Computational Linguistics: ACL 2023*, 11091–11104. Toronto, Canada: Association for Computational Linguistics.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) NA

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) no

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) NA

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) NA

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have

comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes