

# What Makes Math Word Problems Challenging for LLMs?

KV Aditya Srivatsa  
MBZUAI

Abu Dhabi, UAE

vaibhav.kuchibhotla@mbzuai.ac.ae

Ekaterina Kochmar  
MBZUAI

Abu Dhabi, UAE

ekaterina.kochmar@mbzuai.ac.ae

## Abstract

This paper investigates the question of what makes math word problems (MWP) in English challenging for large language models (LLMs). We conduct an in-depth analysis of the key linguistic and mathematical characteristics of MWPs. In addition, we train feature-based classifiers to better understand the impact of each feature on the overall difficulty of MWPs for prominent LLMs and investigate whether this helps predict how well LLMs fare against specific categories of MWPs.<sup>1</sup>

## 1 Introduction

In recent years, large language models (LLMs) have not only demonstrated huge potential across a range of core NLP tasks (Zhao et al., 2023; Brown et al., 2020; Radford et al., 2019, inter alia), but also exhibited a number of emergent abilities, such as an ability to solve mathematical puzzles (Wei et al., 2022). Math word problems (MWPs) have been proposed as a challenging testbed for LLMs, as they test not only the ability of the models to deal with purely mathematical expressions, but also their reasoning and natural language understanding abilities (Wang and Lu, 2023; Cobbe et al., 2021; Patel et al., 2021; Miao et al., 2020, inter alia). Experiments show that even quite powerful LLMs are still challenged by MWPs (Cobbe et al., 2021). At the same time, most previous work has either focused on evaluation of LLMs' performance on MWPs or on changes in their behavior in response to progressive-hint prompting, prompt paraphrasing or similar approaches (Norberg et al., 2023; Raiyan et al., 2023; Zheng et al., 2023; Zhu et al., 2023), while an in-depth analysis of what exactly makes math problems challenging for LLMs is lacking. We aim to address this knowledge gap.

A recent study by Almoubayyed et al. (2023) demonstrates a strong connection between reading

<sup>1</sup>Our code, data, and analysis are publicly available at [github.com/kvadityasrivatsa/analyzing-llms-for-mwps](https://github.com/kvadityasrivatsa/analyzing-llms-for-mwps)

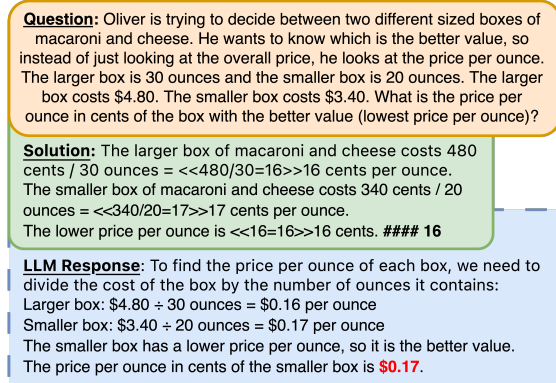


Figure 1: A response from Llama2-70B to a lengthy math problem that involves NLU challenges.

skills and math outcomes in students. We hypothesize that LLMs' ability to solve MWPs correctly may similarly rely on: (1) *the linguistic complexity* of the questions; (2) *the conceptual complexity* of the tasks (e.g., the number of steps and types of math operations involved); and (3) *the amount of real-world knowledge* required to solve the tasks. Supporting this intuition, our preliminary analysis of the GSM8K dataset (Cobbe et al., 2021) suggests that relatively short questions with a small number of described entities, a few calculation steps and a limited range of operators involved in the solution (e.g., *Mark is 7 years older than Amy, who is 15. How old will Mark be in 5 years?*) are typically answered correctly by a range of LLMs. At the same time, long questions requiring real-world knowledge (e.g., how many cents there are in a dollar) and extended natural language understanding (NLU) (e.g., interpretation of a lower price) pose challenges for LLMs (see Figure 1).

In this paper, we formulate and investigate two research questions: (1) *Which characteristics of the input math word questions make them complex for an LLM?* and (2) *Based on these characteristics, can we predict whether a particular LLM will be able to solve specific input MWPs correctly?*

## 2 Methodology

**Data** We use the GSM8K dataset (Cobbe et al., 2021), divided into 7,473 training and 1,319 test instances, because of the high quality of human-generated MWPs. This dataset contains a diverse set of problems in English with minimal amount of recurring templates. Furthermore, the difficulty level of the problems is tailored for LLMs, allowing for a wide variation in correctness across models and question types, which is ideal for our feature-based analysis.

**Approach** We collect solution attempts from several LLMs to the questions from the GSM8K training and test sets. Next, we train statistical classifiers on a filtered subset of questions to predict if they are consistently solved correctly or incorrectly across multiple runs of the models. Our approach is relatively simple but it allows us to investigate which of the features are most indicative of the challenges LLMs face in solving math problems.

**LLMs** We select an array of open-source models for our experiments. We use Llama2 (13B and 70B) (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023) as its performance on math tasks has been found to match models several times its size, and MetaMath-13B (Yu et al., 2023) as it is fine-tuned on math QA data in contrast to the other general-purpose models in the pool.

**Features** We analyze and experiment with the features extracted from MWP questions and their respective expected solutions. This way, the features remain grounded in the dataset, allowing our approach to be applied to any LLM. The features are broadly grouped into the following categories:<sup>2</sup>

1. **Linguistic features** focus on the phrasing of the question. These include the length of the question, sophistication of the vocabulary, syntactic complexity, instances of coreference, and overall readability. Note that the linguistic features are only extracted from the question body as the phrasing of the gold solution has no impact on the expected answer.
2. **Mathematical features** cover the math arguments, operations, and reasoning steps required to solve the questions. These include the number and diversity of the math operations in the solution body. Arguments provided in the question but not utilized in the

<sup>2</sup>The complete list of features extracted, their description and further statistics can be found in Appendix A

| Model        | Success Rate ( $N=1,319$ ) |          |
|--------------|----------------------------|----------|
|              | $\mu$                      | $\sigma$ |
| Llama2-13B   | 0.3724                     | 0.3681   |
| Llama2-70B   | 0.5609                     | 0.3941   |
| Mistral-7B   | 0.3627                     | 0.3309   |
| MetaMath-13B | 0.6373                     | 0.3816   |

Table 1: Success rates for solution attempts per LLM

solution also require mathematical reasoning for them to be disregarded as noise. Note that while a question can be phrased in many ways (affecting its linguistic features), the underlying math operations and reasoning steps (thus, the mathematical features) remain unchanged.

3. **Real-world knowledge & NLU based features** indicate the amount of extraneous information needed to solve the task that is not provided explicitly in the question. This may include how many days there are in a month or the interpretation of “half” as  $1/2$ .

## 3 Experiments

### 3.1 Solution Generation

To collect solution attempts from the LLMs, we use a simple task-specific prompt (See Appendix B) to minimize any bias imposed on the model generation. We query each LLM 5 times on each question with varying generation seeds and a temperature of 0.8. A soft-matching strategy is then used to extract the final answer from the solutions. Using each LLM’s attempted solutions, every question is assigned a mean **success rate** using  $(\# \text{ of correct answers}) / (\# \text{ of solution attempts})$ .

### 3.2 Success Rate Prediction

We train and evaluate classifiers on their ability to predict for input test questions whether they will be answered correctly or incorrectly by a specific LLM. We also train and evaluate classifiers on the **intersection** set of questions, which are either solved correctly by *all* or by *none* of the LLMs.

**Models** We use Logistic Regression, Decision Tree, and Random Forest classifiers, which allow us to extract relative feature importance with ease.

**Data** For high confidence samples, we use the training and test subset from GSM8K where the sampled success rate is either 1.0 (**always** correct) or

0.0 (**never** correct). The distribution of the LLM-specific splits is detailed in Table 2.

**Preprocessing & Optimization** We employ several preprocessing steps including dropping highly correlated features, class-balancing, and feature scaling. We also perform a hyperparameter search for each model to maximize performance on unseen data. See Appendix C for more details.

## 4 Results

### 4.1 Success Rate Distribution

We report the mean success rates for each LLM on GSM8K’s test set in Table 1.<sup>3</sup> We observe that Llama2 13B and 70B follow the expected order of scores along their respective parameter counts. Mistral-7B scores similar to the 13B Llama2 model, and the additional fine-tuning allows MetaMath-13B to outperform the other models (including the 70B Llama2). Figures 2a and 2b respectively capture the number of questions *always* and *never* answered correctly by each LLM. Overall, MetaMath-13B has the lowest number of incorrectly and the highest number of correctly answered questions across the tested LLMs.

### 4.2 Classification Results

To compare classifiers’ performance, we report the accuracy and macro-F1 scores for each classifier and LLM-specific test data split (see Table 2). We observe that Random Forest outperforms other classifiers across most solution sets.

At the same time, we also note that, due to significant class imbalance, this task is not easy for the classifiers, with the best accuracy scores across LLM splits being in the range of 71.7% – 81.4%. The small number of questions always or never solved correctly by any LLM speaks to the models’ varying capabilities (and potential points of brittleness). We include additional analysis of the results in Appendix D.

For comparison, we also report the classification results for a fine-tuned RoBERTa-base model (Liu et al., 2019) for the same training and evaluation sets (tuned on the question and gold solution as input text; see Appendix C for more details) in Table 2. We note that the Transformer base classifier scores on a par or a few points above the best statistical classifier, i.e., Random Forest, suggesting

<sup>3</sup>Our results generally align with those reported previously for these models.

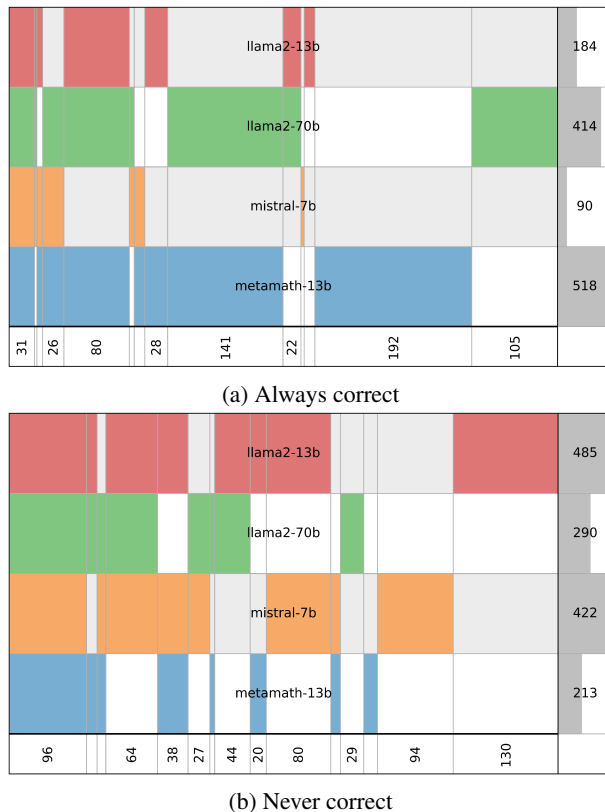


Figure 2: Number of questions from the GSM8K-test (a) always and (b) never answered correctly by each LLM. The rows in each figure correspond to individual LLMs, with the counts on the right denoting the total number of questions always (or never) answered correctly by each LLM. The counts at the bottom denote the number of questions in each subset of LLMs.

that the proposed feature-based classifiers are not far behind token-level contextual models for this task.

#### 4.2.1 Feature Importance

The statistical classifiers used in our experiments allow us to estimate the importance of each feature and its contribution to the classification performance. We report the top 10 features with the highest aggregate ranks across LLM data splits and classifiers in Table 3. We use mean rank here as a proxy for relative importance across features, and the respective standard deviations indicate how spread out this importance is across classifiers and queried LLMs. We observe that a greater number (Gx\_op\_unique\_count) and diversity (Gx\_op\_diversity) in math operations, and the use of infrequent numerical tokens in the question and solution body (Qx\_ & Gx\_mean\_numerical\_word\_rank) impact the success rate. The list also contains linguistic fea-

| Class Distribution |        |               |               |               |              |              |
|--------------------|--------|---------------|---------------|---------------|--------------|--------------|
| Split              | Class  | Llama2-13b    | Llama2-70b    | Mistral-7b    | MetaMath-13b | Intersection |
| Train              | Always | 1102 (30.22%) | 2438 (61.36%) | 733 (24.06%)  | 5162 (94.7%) | 205 (53.38%) |
|                    | Never  | 2545 (69.78%) | 1535 (38.64%) | 2313 (75.94%) | 289 (5.3%)   | 179 (46.61%) |
|                    | Total  | 3647          | 3973          | 3046          | 5451         | 401          |
| Test               | Always | 188 (28.14%)  | 427 (60.06%)  | 111 (21.51%)  | 528 (71.64%) | 31 (24.41%)  |
|                    | Never  | 480 (71.86%)  | 284 (39.94%)  | 405 (78.49%)  | 209 (28.36%) | 96 (75.59%)  |
|                    | Total  | 668           | 711           | 516           | 737          | 135          |

| Classification Performance |              |              |              |              |              |              |              |              |              |              |
|----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Classification Model       | Llama2-13b   |              | Llama2-70b   |              | Mistral-7b   |              | MetaMath-13b |              | Intersection |              |
|                            | Acc.         | Macro F1     | Acc.         | Macro F1     | Acc.         | Macro F1     | Acc.         | Macro F1     | Acc.         | Macro F1     |
| Logistic Regression        | 0.707        | 0.686        | 0.684        | 0.673        | 0.721        | 0.675        | 0.737        | <b>0.686</b> | 0.800        | 0.787        |
| Decision Tree              | 0.657        | 0.625        | 0.644        | 0.637        | 0.667        | 0.611        | 0.703        | 0.627        | 0.733        | 0.719        |
| Random Forest              | <b>0.767</b> | <b>0.724</b> | <b>0.717</b> | <b>0.707</b> | <b>0.814</b> | <b>0.738</b> | <b>0.744</b> | 0.549        | <b>0.815</b> | <b>0.799</b> |
| RoBERTa-base               | 0.816        | 0.771        | 0.756        | 0.738        | 0.838        | 0.743        | 0.701        | 0.415        | 0.811        | 0.781        |

Table 2: Class-wise distribution and classification results for different LLMs. "Intersection" refers to questions always or never solved correctly by all or any LLM, respectively. All classification results are mean values across 5 runs with varying initialization seeds. The best results for feature-based classifiers are highlighted in bold.

| Type | Feature Name                | Rank ( $N=23$ ) |          |
|------|-----------------------------|-----------------|----------|
|      |                             | $\mu$           | $\sigma$ |
| L    | Qx_np_count                 | 1.2             | 0.45     |
| M    | Qx_mean_numerical_word_rank | 4               | 1.87     |
| M    | Gx_op_unique_count          | 4               | 2.65     |
| M    | Gx_op_diversity             | 4.4             | 2.30     |
| M    | Gx_mean_numerical_word_rank | 4.4             | 1.82     |
| L    | Qx_mean_word_rank           | 5.6             | 1.82     |
| L    | Qx_flesch_kinkaid_grade     | 6               | 1.87     |
| W    | Gx_world_knowledge          | 7.8             | 2.28     |
| L    | Qx_constituency_tree_depth  | 9.6             | 1.95     |
| M    | Gx_op'+'_count              | 11.6            | 3.97     |

Table 3: Feature importance ranks across classification models and LLM-wise data subsets.

tures based on the phrasing of the questions: longer questions with a high number of noun phrases (Qx\_np\_count), mean syntactic depth (Qx\_constituency\_tree\_depth), and readability grade (Qx\_flesch\_kinkaid\_grade) are also difficult for LLMs to solve. Additionally, the need for extraneous information (Gx\_world\_knowledge), such as conversion units for time, distance, or weight, can make a question challenging. We also report value thresholds at which each feature affects the success rate significantly: see the results of the Student's  $t$ -test and  $p$ -values in Table 7 in Appendix D.

#### 4.2.2 Ablation Studies

To further measure the impact of each feature type, we report classification scores along different feature-type subsets in Figure 3. We note that the feature set with all types (L+M+W) is not optimal for

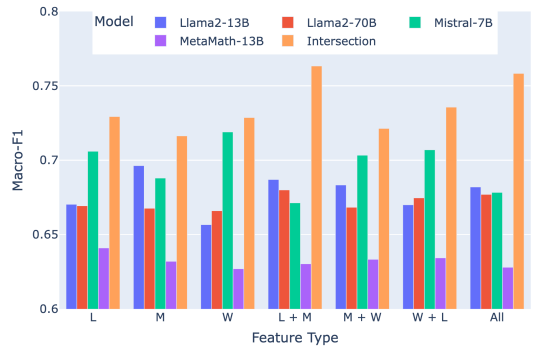


Figure 3: Results of the ablation studies across feature types (L – Linguistic, M – Mathematical, W – World Knowledge & NLU). Each bar represents the mean macro-F1 score over all three classifier models.

classification. For instance, the questions answered by Llama2-13B are best classified using only mathematical features (M). The best-performing classifiers for Llama2-7B, MetaMath-13B, and the intersection set either solely use linguistic features (L) or both linguistic and math features (L+M), whereas the world knowledge & NLU feature set is sufficient for Mistral-7B.

#### 4.2.3 Impact of Linguistic Features

In order to better gauge the impact of linguistic features on the success rate, we cluster questions by mathematical features. We fit a KMeans clustering model<sup>4</sup> on all math features for each question in the GSM8K training set with a target cluster count of 100. This helps group together questions from the data, wherein the math features hardly vary within each question subset (or cluster). Thus, vari-

<sup>4</sup><https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

| Cluster |      | Feature                    | Spearman( $\rho$ ) |
|---------|------|----------------------------|--------------------|
| ID      | Size |                            |                    |
| 09      | 27   | Qx_constituency_tree_depth | -0.64***           |
| 24      | 14   | Qx_mean_word_rank          | -0.82***           |
| 63      | 62   | Qx_token_length            | -0.42***           |
|         |      | Qx_word_length             | -0.43***           |
| 96      | 51   | Qx_flesch_kinkaid_grade    | -0.51***           |

Table 4: Cluster-wise feature correlations. The cluster count represents the number of questions included in the respective cluster. The p-value for all reported correlation values is  $<0.001$  (marked by ‘\*\*\*’).

ations in success rate across the questions within a cluster can be more clearly attributed to other, i.e., linguistic types of features. We report some notable Spearman correlation values between the linguistic feature values within a cluster and the corresponding success rates in Table 4. The strong and significant feature-wise negative correlations suggest that for a relatively fixed set of math features, questions with greater length, nesting, lexical rank, and reading grade become more challenging for LLMs to solve. Note that this form of analysis on feature-based minimal pairs is extractive in nature and may, to a certain extent, be restricted to the question types in the GSM8K dataset. For a more exhaustive analysis for each feature, generative approaches to furnish question paraphrases with the desired set of linguistic features need to be employed.

## 5 Conclusions

This work aims to identify what aspects of MWP make them difficult for LLMs to solve. To this end, we extract key features (spanning linguistic, mathematical, and real-world knowledge & NLU-based aspects) to predict whether several LLMs can reliably solve MWPs from GSM8K. We find that questions with a high number and diversity of math operations using infrequent numerical tokens are particularly challenging to solve. In addition, we show that lengthy questions with low readability scores and those requiring real-world knowledge are also seldom solved correctly. Our future work will rely on these findings to make informed modifications to questions in order to study the impact on LLMs’ reasoning and MWP-solving abilities. Figure 4 provides an example of an informed modification, which leads to improved LLM performance.

**Question:** Oliver is trying to decide between two different sized boxes of macaroni and cheese. He wants to know which is the better value, so instead of just looking at the overall price, he looks at the price per ounce. The larger box is 30 ounces and the smaller box is 20 ounces. The larger box costs \$4.80. The smaller box costs \$3.40. What is the price per ounce in cents of the box with the better value (lowest price per ounce)? **Note that \$1 is 100 cents.**

**LLM Response:** To find the price per ounce in cents of the box with the better value, we need to divide the cost of each box by the number of ounces it contains. For the larger box: Price per ounce =  $\$4.80 \div 30 \text{ ounces} = 16 \text{ cents per ounce}$ . For the smaller box: Price per ounce =  $\$3.40 \div 20 \text{ ounces} = 17 \text{ cents per ounce}$ . Since 16 cents per ounce is less than 17 cents per ounce, the larger box has the better value. So, the price per ounce in cents of the box with the better value (lowest price per ounce) is **16 cents** per ounce.

Figure 4: Solution attempt by Llama2-70B on the question from Figure 1, with the required real-world knowledge explicitly specified.

## Limitations

With the rapidly growing body of research on LLMs, this study necessarily has several limitations, which we discuss below.

**Limited set of LLMs tested** We consider it important to test and report results with a diverse set of open-source LLMs, which motivated the selection of the specific models included in this study. At the same time, we do not claim this study to be comprehensive with respect to the range of LLMs tested and in future work, we plan to include more LLMs in this research.

**Limited number of classification models** As the main goal of this study is to identify aspects of the MWPs that make them challenging for LLMs to solve, we have opted for a feature-based approach and a range of traditional classification models as opposed to less transparent but more powerful black-box algorithms. Our results show that the prediction task is challenging for the traditional classifiers that we used, and it is likely that these results can be improved with stronger classification models.

**Limitations of the dataset** In this work, we have focused on a single MWP dataset (GSM8K) due to its unique properties, namely the high quality of the questions, high diversity of the tasks (including linguistic diversity of the questions), and moderate difficulty of the math problems covered (Cobbe et al., 2021). At the same time, we recognize that the results we report in this work may be limited in certain ways to the dataset on which we report them. Our future work will apply this approach to other available MWP datasets (Kim et al., 2023; Wang and Lu, 2023; Patel et al., 2021; Miao et al., 2020) to verify the consistency of the findings.

**Impact studies** Finally, whilst we have identified aspects of the MWP that make them challenging for LLMs to solve, we admittedly presented only one example (see Figure 4) where acting upon one of the identified aspects improves the output of an LLM. While a thorough investigation of the impact of such informed modifications is outside the scope of the current paper, such experiments will follow in future work to demonstrate the practical usefulness of the identified MWP aspects.

## Ethics Statement

We foresee no serious ethical implications from this study.

## Acknowledgments

We are grateful to the Campus Super Computing Center at MBZUAI for supporting this work. We also thank the anonymous reviewers for their valuable feedback.

## References

- Husni Almoubayyed, Stephen E Fancsali, and Steve Ritter. 2023. Instruction-embedded assessment for reading ability in adaptive mathematics software. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 366–377.
- Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*.
- 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized BERT pretraining approach*. *CoRR*, abs/1907.11692.
- 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.
- Kole Norberg, Husni Almoubayyed, Stephen E. Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steven Ritter. 2023. *Rewriting math word problems with large language models*. In *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), Tokyo, Japan, July 7, 2023*, volume 3487 of *CEUR Workshop Proceedings*, pages 163–172. CEUR-WS.org.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. *Are NLP models really able to solve simple math word problems?* In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Syed Rifat Raiyan, Md Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. *Math word problem solving by generating linguistic variants of problem statements*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 362–378, Toronto, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Tianduo Wang and Wei Lu. 2023. Learning multi-step reasoning by solving arithmetic tasks. In *Proceedings of ACL*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. [Math: Bootstrap your own mathematical questions for large language models](#).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

## A Feature Details

Below, we describe the features used in our study and how they were extracted.

(1) **Linguistic features (L)** include 9 features pertaining to the question (Q) itself:

- **Qx\_token\_length**: The number of tokens in the tokenized version of the question body. We apply each LLM’s respective tokenizer from [HuggingFace](#) to extract this feature.
- **Qx\_sentence\_length**: The number of sentences detected in the question body. We use the [sentence\\_splitter](#) Python library to extract this count.
- **Qx\_word\_length**: The number of space-separated segments (words) in the question body.
- **Qx\_flesch\_kincaid\_grade**: The readability grade of the question body as per the FKGL metric ([Flesch, 1948](#)). We use the [textstat](#) Python library to extract this feature.
- **Qx\_mean\_word\_rank**: The mean vocabulary rank (in decreasing order of frequency) of the tokens in the question body. We use the same tokenizer set used for `Qx_token_length`.
- **Qx\_constituency\_tree\_depth**: The mean depth of the constituency tree across the sentences in the question body. We use Stanford’s [Stanza](#) parsing library to parse the questions.
- **Qx\_np\_count**: Number of distinct noun phrases detected in the question body. We extract this from the constituency parse collected from the [Stanza](#) parser.
- **Qx\_prp\_count**: Number of prepositions in the question body. We use the part-of-speech tags generated as part of the parse by [Stanza](#).
- **Qx\_coref\_count**: Number of pronominal or nominal instances of coreference in the question body. We use Stanford’s [CorefAnnotator](#) to extract this feature.

(2) **Mathematical features (M)** include 12 features pertaining to the question (Q) and gold solution (G):

| Type | Source | #                  | Feature Name                | Range                | $\mu$     | $\sigma$ |
|------|--------|--------------------|-----------------------------|----------------------|-----------|----------|
| L    | Q      | 1                  | Qx_token_length             | [12 – 239]           | 66.05     | 24.384   |
|      | Q      | 2                  | Qx_sentence_length          | [1 – 13]             | 3.431     | 1.201    |
|      | Q      | 3                  | Qx_word_length              | [9 – 184]            | 45.885    | 17.832   |
|      | Q      | 4                  | Qx_flesch_kincaid_grade     | [-1.9 – 26.3]        | 4.236     | 2.468    |
|      | Q      | 5                  | Qx_mean_word_rank           | [3661.96 – 21929.96] | 10646.615 | 2110.891 |
|      | Q      | 6                  | Qx_constituency_tree_depth  | [5 – 31]             | 10.803    | 2.798    |
|      | Q      | 7                  | Qx_np_count                 | [3 – 74]             | 18.034    | 7.488    |
|      | Q      | 8                  | Qx_prp_count                | [0 – 16]             | 1.772     | 1.854    |
|      | Q      | 9                  | Qx_coref_count              | [0 – 16]             | 0.462     | 1.283    |
|      | Q      | 10                 | Qx_arg_count                | [0 – 17]             | 4.438     | 1.94     |
|      | Q      | 11                 | Qx_word_arg_count           | [0 – 14]             | 1.091     | 1.397    |
| M    | Q      | 12                 | Qx_mean_numerical_word_rank | [259.0 – 29905.38]   | 22643.319 | 3260.09  |
|      | G      | 13                 | Gx_arg_count                | [6 – 73]             | 24.377    | 9.732    |
|      | G      | 14                 | Gx_op‘+’_count              | [0 – 12]             | 1.06      | 1.212    |
|      | G      | 15                 | Gx_op‘-’_count              | [0 – 6]              | 0.601     | 0.78     |
|      | G      | 16                 | Gx_op‘*’_count              | [0 – 8]              | 1.369     | 1.183    |
|      | G      | 17                 | Gx_op‘/’_count              | [0 – 7]              | 0.621     | 0.789    |
|      | G      | 18                 | Gx_op‘(’_count              | [0 – 4]              | 0.026     | 0.187    |
|      | G      | 19                 | Gx_op_unique_count          | [0 – 6]              | 2.284     | 0.93     |
|      | G      | 20                 | Gx_op_diversity             | [0.15 – 1.0]         | 0.758     | 0.196    |
|      | G      | 21                 | Gx_mean_numerical_word_rank | [22645.0 – 29915.0]  | 28626.04  | 776.73   |
| B    | 22     | Gx_parameter_usage | [0.07 – 1.0]                | 0.642                | 0.241     |          |
| W    | B      | 23                 | Gx_world_knowledge          | [0 – 8]              | 1.104     | 1.006    |

Table 5: Details of formulation and distribution (across GSM8K) for all features included in the feature set. Each feature is of type: Linguistic (L), Mathematical (M), or World Knowledge and NLU (W) and is sourced either from the question body (Q), gold solution body (G), or both (B).

- **Qx\_arg\_count:** The number of distinct numerical quantities (e.g., “**3.5** hours later” or “**100** boxes”) in the question body. We use a Regexp pattern to detect whole numbers, decimal point values, and quantities preceded by a negative sign or dollar (and other currency) signs.
- **Qx\_word\_arg\_count:** The number of quantities mentioned in word-form (“**three** times” or “**half** as much”) in the question body. We use a vocabulary of frequently used word-form tokens and accommodate compound expressions (e.g., “twenty-two”).
- **Qx\_mean\_numerical\_word\_rank:** The mean vocabulary rank of the numerical tokens in the question body. We first isolate numerical tokens tokenized by respective tokenizers, then aggregate their token rank.
- **Gx\_arg\_count:** The number of distinct numerical quantities present as plain text or on the left-hand side of equations in the gold solution. We use the same Regexp pattern used for Qx\_arg\_count.
- **Gx\_op{‘+’/‘-’/‘\*’/‘/’/‘(’}\_count:** Number of times each listed math operation is used in the gold solution. A simple Regexp pattern is applied to extract these from within equations.
- **Gx\_op\_unique\_count:** The maximum number of times a single operation has been used in the gold solution. For instance, “ $3 + 4.5 + 7 + 1 - 2.7$ ” contains 3 instances of the ‘+’ operator.
- **Gx\_op\_diversity:** Ratio of the number of unique math operators used to the total number of operators in the gold solution. For instance, a question with the consolidated math solution expression “ $(2 \times 12) \times 3 = 72$ ” contains two arithmetic operations in total but only one unique operation type, i.e., ‘ $\times$ ’,  $Gx\_op\_diversity = 1/2 = 0.5$ .
- **Gx\_mean\_numerical\_word\_rank:** The mean vocabulary rank of the numerical tokens used on the left-hand side of equations in the



gold solution. Extracted the same way as `Qx_mean_numerical_word_rank`.

- **Bx\_parameter\_usage**: The ratio of distinct arguments used in the gold solution to that in the question body. A value lower than 1 indicates that one or more arguments provided in the question were not required to solve the MWP (potentially acting as distractors).

(3) **World knowledge and NLU features (W)** include:

- **Bx\_world\_knowledge**: The number of distinct arguments on the left-hand-side of equations in the gold solution, that are neither present in the question body nor produced as intermediate results from any prior equations in the solution. A non-zero value is interpreted as the use of a quantity (perhaps a conversion factor, or the number of entities involved in computing mean) unspecified by the question. The arguments were extracted from both sides using the same Regexp policies used for previous features.

Table 5 shows further statistics on the features, including the range as well as the mean and standard deviation of the values for each feature type. Additionally, we report the Spearman correlation between all pairs in the feature set in Figure 5.

## B Querying Details

### B.1 Prompt Template

We use a simple task-specific prompt (see Figure 6) either prepended to the question-prompt or specified as a system-prompt if an LLM input query format requires so.

### B.2 LLM Details

The exact large language models used in our experiments, along with their reported performance on GSM8K according to the OpenLLM leaderboard (Beeching et al., 2023) are mentioned in Table 6. All LLMs and libraries used are open-source. The license to use Meta’s Llama2 models was procured through due process.

| Model        | HuggingFace Model Name             | Pass@1 |
|--------------|------------------------------------|--------|
| Llama2-13B   | meta-llama/Llama-2-13b-chat-hf     | 28.70  |
| Llama2-70B   | meta-llama/Llama-2-70b-chat-hf     | 56.80  |
| Mistral-7B   | mistralai/Mistral-7B-Instruct-v0.2 | 40.03  |
| MetaMath-13B | meta-math/MetaMath-13B-V1.0        | 72.30  |

Table 6: List of HuggingFace model variants and their respective reported pass@1 (single run) accuracies on the GSM8K test set from the OpenLLM leaderboard.

### B.3 Implementation and Compute Resources Used

We use `vLLM` to load and query models. Models of parameter sizes 7B and 13B were queried with a single NVIDIA A100 GPU. Llama2-70B was loaded and queried using 4 A100 GPUs. Each query was set to a temperature of 0.8, and a maximum token length of 2000. Each question was queried 5 times by each LLM, with a varying seed. Querying the entire GSM8K dataset (8,793 questions) took approximately 1 hour for each LLM.

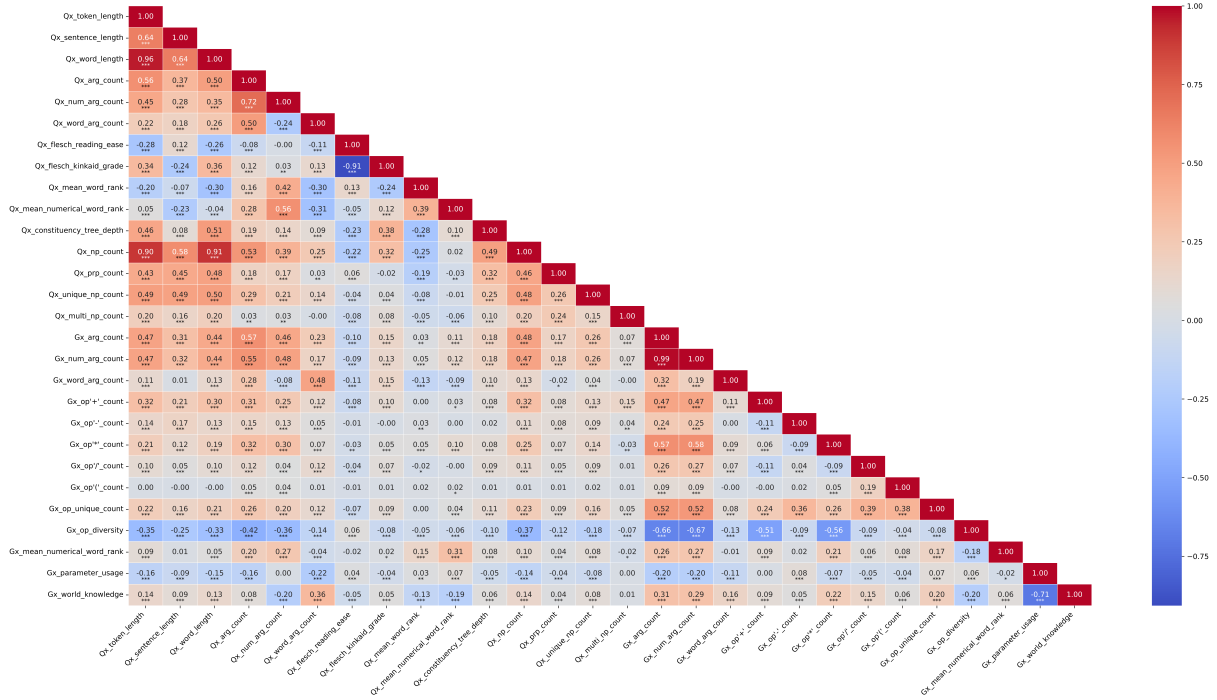


Figure 5: Spearman correlation matrix between features. All correlation values are marked with ‘\*’, ‘\*\*’, and ‘\*\*\*’ if their corresponding p-values are less than 0.05, 0.01, and 0.001 respectively.

| Feature            | Llama2-13B |       |        | Llama2-70B |       |        | Mistral-7B |       |         | MetaMath-13B |       |         | Intersection |       |         |
|--------------------|------------|-------|--------|------------|-------|--------|------------|-------|---------|--------------|-------|---------|--------------|-------|---------|
|                    | Thresh.    | Diff. | T-val  | Thresh.    | Diff. | T-val  | Thresh.    | Diff. | T-val   | Thresh.      | Diff. | T-val   | Thresh.      | Diff. | T-val   |
| Gx_num_arg_count   | 51.102     | 0.313 | -8.328 | 53.273     | 0.380 | -7.747 | 55.082     | 0.283 | -5.452  | 59.061       | 0.438 | -5.558  | 57.735       | 0.330 | -5.977  |
| Qx_np_count        | 40.673     | 0.200 | -5.440 | 47.465     | 0.435 | -5.090 | 33.429     | 0.207 | -11.201 | 47.918       | 0.448 | -7.230  | 47.918       | 0.333 | -5.645  |
| Gx_arg_count       | 57.959     | 0.336 | -6.018 | 58.111     | 0.380 | -5.742 | 57.959     | 0.289 | -5.644  | 51.122       | 0.271 | -10.323 | 60.694       | 0.335 | -5.249  |
| Qx_word_length     | 91.143     | 0.219 | -7.944 | 97.384     | 0.300 | -8.468 | 19.714     | 0.233 | -7.016  | 101.857      | 0.275 | -9.267  | 116.143      | 0.251 | -5.328  |
| Gx_op*+_count      | 2.082      | 0.198 | -7.683 | 2.061      | 0.246 | -9.009 | 2.082      | 0.181 | -7.660  | 3.061        | 0.339 | -7.348  | 3.061        | 0.280 | -6.383  |
| Qx_prp_count       | 7.184      | 0.215 | -5.957 | 8.081      | 0.313 | -6.322 | 8.163      | 0.235 | -5.475  | 7.184        | 0.167 | -6.039  | 8.163        | 0.233 | -6.834  |
| Qx_sentence_length | 6.143      | 0.206 | -6.843 | 6.091      | 0.264 | -8.278 | 6.143      | 0.203 | -7.357  | 7.122        | 0.253 | -5.756  | 7.122        | 0.229 | -5.473  |
| Gx_op*+_count      | 4.163      | 0.241 | -7.849 | 4.121      | 0.293 | -8.996 | 5.143      | 0.296 | -6.038  | 3.184        | 0.080 | -5.377  | 5.143        | 0.238 | -6.081  |
| Qx_token_length    | 123.184    | 0.211 | -8.937 | 126.646    | 0.258 | -9.557 | 113.918    | 0.214 | -12.560 | 132.449      | 0.232 | -9.953  | 118.551      | 0.219 | -14.125 |
| Gx_op**+_count     | 4.082      | 0.211 | -6.135 | 4.040      | 0.220 | -6.043 | 4.082      | 0.184 | -5.831  | 5.061        | 0.268 | -5.423  | 5.061        | 0.250 | -5.295  |

Table 7: Feature-wise thresholds which reflect the greatest difference in the corresponding mean success rate. For each feature, the optimal threshold creates two sets of questions on either side, wherein the difference in the corresponding mean success rates of the two sets is the greatest. We perform Student’s *t*-tests on both sets to determine if this difference is significant and report the corresponding *t* values. All results reported in the table have an absolute *t*-value >5 and a *p*-value <0.0001.

You are an expert in solving math questions.  
 Answer the following question to the best of  
 your ability.  
 <<Question\_Body>>

Figure 6: Prompt template used for solution generation across LLMs.

## C Training Details

### C.1 Preprocessing

Before training classifiers, we perform the following steps on the feature data:

1. **Pruning:** For each training feature-set, we

iteratively remove features with high correlation with other features until no two columns in the data have an absolute Spearman correlation higher than 0.5.

2. **Scaling:** We fit scikit-learn’s `StandardScaler` onto the train split to normalize the mean and standard deviation of all features. We then apply the same scaler on the test split features.
3. **Balancing:** As many LLM solution splits either contain too few always-correct or always-incorrect question samples, we use imblearn’s `RandomOverSampling` tool to balance the pro-

| Classification Model | Llama2-13B    |               | Llama2-70B    |               | Mistral-7B    |               | MetaMath-13B  |               | Intersection  |               |
|----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                      | F1(Never)     | F1(Always)    | F1(Never)     | F1(Always)    | F1(Never)     | F1(Always)    | F1(Never)     | F1(Always)    | F1(Never)     | F1(Always)    |
| Logistic Regression  | 0.8454        | 0.4968        | 0.5641        | <b>0.7862</b> | 0.8926        | 0.5081        | 0.0095        | 0.8354        | 0.8263        | <b>0.7184</b> |
| Decision Tree        | 0.8199        | 0.4532        | 0.4958        | 0.7431        | 0.8732        | 0.4592        | <b>0.0727</b> | <b>0.8373</b> | 0.732         | 0.6496        |
| Random Forest        | <b>0.8535</b> | <b>0.5049</b> | <b>0.6133</b> | 0.7824        | <b>0.8936</b> | <b>0.5161</b> | 0.019         | 0.8361        | <b>0.8439</b> | 0.7216        |

Table 8: Class-wise F1 scores for each classification model across LLM solution splits.

portions of the two classes in each run.

## C.2 Hyperparameter Search

For each classifier model and LLM solution split pair, we conduct Bayesian optimization on the ranges of key hyperparameters for each classifier. As the objective function, we maximize the Macro-F1 score on a 15% held-out set of GSM8K’s test split to prevent the models from over-fitting onto the train samples.

## C.3 Fine-tuning RoBERTa Classifier

We use HuggingFace’s [Trainer](#) module to tune a pre-trained [RoBERTa-base](#) classifier on the same target data as the statistical classifiers in each setting. The corresponding input text for training and evaluation was built by concatenating the GSM8K question and gold-solution text for each sample, i.e., "`<question> Question Body </question>`" `<solution> Solution Body </solution>`". The model is trained for 3 epochs with a peak learning rate of  $2e - 5$  and a warmup ratio of 0.1. On a Tesla P100, each training run took approximately 10-15 minutes.

## D Results Analysis

### D.1 Class-wise Classification Review

Table 8 reports the class-wise F1-scores (for questions that are always or never solved correctly) for each classifier across LLM question splits. Though there may be notable class imbalances among the two classes across splits, all classifiers were trained with proportional oversampling. We see that for relatively smaller pretrained models, i.e., Llama2-13B and Mistral-7B, F1 scores for always-incorrect questions are significantly higher than their counterparts. Thus, the scores indicate that for smaller models, questions answered incorrectly are more predictable. For larger models like Llama2-70B, this difference is lower, with the always-correct questions being somewhat more predictable. For the fine-tuned MetaMath-13B model, the small number of questions that are never answered correctly, fail to provide a generalizable sample for predicting on unseen data.

### D.2 Feature Impact

We continue our discussion of feature importance (from Section 4.2.1) by identifying pivot points for key features about which the corresponding success rates for questions show a significant difference in mean values. We perform Student’s *t*-tests on equally spaced thresholds along each feature and report the thresholds which show the highest variation in mean success rates in Table 7.

We see that across most LLMs, a significant rise in the mean success rate is observed as the question contains, on average, more than 6-7 sentences, 90-115 words, or 113-132 tokens. We get a better idea of the kind of questions Mistral-7B gets wrong more often than other models, as its threshold (19.71 words) for the number of words in the question body is substantially lower than the average. Across features involving number of math operations, the threshold for the fine-tuned MetaMath-13B model is either on a par or higher than other models.