CUTTING THROUGH THE NOISE: BOOSTING LLM PERFORMANCE ON MATH WORD PROBLEMS

Ujjwala Anantheswaran [1◇*] **Himanshu Gupta** [1‡*] **Kevin Scaria** [1‡*] **Shreyas Verma**[2]
Chitta Baral[1] **Swaroop Mishra**[1†]
[1]Arizona State University [2]Georgia Institute of Technology
{uananthe, hgupta35, kscaria}@asu.edu

ABSTRACT

Large Language Models (LLMs) excel at various tasks, including solving math word problems (MWPs), but struggle with real-world problems containing irrelevant information. To address this, we propose a prompting framework that generates adversarial variants of MWPs by adding irrelevant variables. We introduce a dataset, PROBLEMATHIC containing both adversarial and non-adversarial MWPs. Our experiments reveal that LLMs are susceptible to distraction by numerical noise, resulting in an average relative performance drop of $\sim 26\%$ on adversarial MWPs. To mitigate this, we fine-tune LLMs (Qwen-2, Mistral) on the adversarial samples from our dataset. Fine-tuning on adversarial training instances improves performance on adversarial MWPs by $\sim 8\%$, indicating increased robustness to noise and improved ability to identify relevant data for reasoning. Finally, to assess the generalizability of our prompting framework, we introduce GSM-8K-Adv, an adversarial variant of the GSM-8K benchmark. LLMs continue to struggle when faced with adversarial information, reducing performance by up to 24%.

1 INTRODUCTION

Large Language Models (LLMs) (AI@Meta, 2024; Team et al., 2023; Achiam et al., 2023) have achieved impressive performance on various tasks, including mathematical problem solving (Imani et al., 2023; Gaur & Saunshi, 2023; Romera-Paredes et al., 2023; Ahn et al., 2024). Existing mathematical datasets (Patel et al., 2021; Sawada et al., 2023; Mishra et al., 2022; Gupta et al., 2023) typically feature simplified questions with limited variables and numerical data directly relevant to the problem. However, real-world MWPs arise within the context of broader reasoning tasks. Thus, they are likely to contain irrelevant information stemming from the wider text. This unrelated information acts as noise, distracting the reasoning process of language models (Fig. 1). Recent works have shown that LLMs do not perform consistently well on mathematical reasoning (Ahn et al., 2024; Shakarian et al., 2023; Chang et al., 2023) and are sensitive to linguistic variations in MWPs (Patel et al., 2021; Shi et al., 2023; Kumar et al., 2021). Previous works (Kumar et al., 2022) have attempted to make LLMs more robust by generating "adversarial" training data but fail to emulate real-world data by not introducing textual noise.

To initiate a systematic study, we introduce a new dataset PROBLEMATHIC containing both adversarial and non-adversarial mathematical word problem (MWP) pairs. The dataset includes problems of two difficulty levels: *Simple* and *Complex*. These adversarial MWPs are augmented to add irrelevant numerical information while maintaining the integrity of the original question; i.e. their solution and final answer are the same as their non-adversarial counterparts. We employ various LLMs, including Gemini-1.5 Pro, Claude-3.5 Sonnet, and Qwen-1.5, among others, for inference. Our observations reveal significant performance drops (16.07%, 22.93%, 41.53%, respectively) between adversarial and non-adversarial problems, highlighting the models' inability to reason accurately with noisy data. To address this challenge, we propose a prompting framework that allows us to leverage existing MWPs of varying complexities and generate augmented adversarial versions for the training set. We demonstrate that fine-tuning LLMs on these samples enhances their ability to identify irrelevant variables. Fine-tuning Llama-2 (7B, 13B) and Mistral (7B) on these samples yields an average 8% improvement in performance on both Simple (8.19%) and Complex (7.07%) adversarial test samples. Finally, to assess the generalizability of our prompting framework, we create an adversarial variant of the widely used GSM-8K benchmark (GSM-8K-Adv). This reveals that LLMs (Mistral Large, Llama-2, and Llama-3) experience an average performance drop of $\sim 4\%$ when tested on this adversarial benchmark. Additionally, we compare our approach with others in the domain of adversarial data generation in §4.

Contributions: (a) We introduce PROBLEMATHIC, a dataset of *Simple* and *Complex* MWPs that demonstrates the susceptibility of LLMs to irrelevant numerical information. (b) We propose a prompting framework to generate adversarial variants of existing MWPs. We show that fine-tuning on these adversarial samples leads to improved performance and robustness to noise. (c) Using this prompting framework, we introduce GSM-8K-Adv, an adversarial variant of the GSM-8K benchmark.

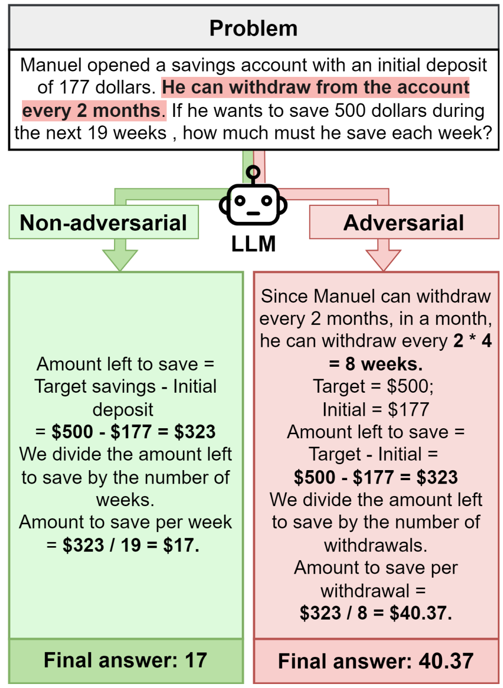

Figure 1: An example showing how adding noise (highlighted in red) to an MWP results in incorrect reasoning and an inaccurate solution. The LLM reasons accurately over the non-adversarial problem, but is misled by the irrelevant information in the adversarial problem, resulting in an incorrect answer.

2 METHODOLOGY

2.1 ADVERSARIAL DATA GENERATION

We propose a constrained additive approach, wherein we add numerical information (*Structural Invariance*) to the original MWP. We use our hand-crafted set of MWPs as seed samples and augment them in accordance with the following constraints: (1) The added variables must not be related to or derived from the existing variables in the passage. (2) The added variables must not share the same physical unit as any of the original variables. (3) The augmented text should not add any numerical information about existing variables that did not exist in the original passage. This gives us a set of adversarial MWPs containing variables that do not feature in their solution equations. Hence, this allows us to evaluate the model's ability to identify relevant information and demonstrate reasoning ability beyond logically connecting all numerical entities in the input in order to produce a solution. The prompting framework used to adversarially augment the seed samples can be found in §C. We also compare our adversarial data generation with other contemporaneous approaches (§4).

2.2 DATASET

PROBLEMATHIC is split into *Simple* and *Complex* problem sets, containing both adversarial and non-adversarial samples. Building upon prior efforts in mathematical problem dataset curation, we leverage existing MWP datasets (Koncel-Kedziorski et al., 2016; Roy & Roth, 2016; 2018; Kushman et al., 2014). These MWPs are aggregated, undergo data cleaning procedures, and are subsequently categorized into distinct *Simple* and *Complex* subsets based on the number and complexity of mathematical operations required to reach the solution. We distinguish between *Simple* and *Complex* problems as follows:

- *Simple*: MWPs involving no more than 3 linear operations. These problems require addition and subtraction operations only. Number of samples in train and test splits are 1313 and 60 respectively.
- *Complex*: MWPs involving multiple nested operations. These questions often involve rates and proportions. They require multiplication and division operations too. Number of samples in train and test splits are 613 and 73 respectively.

Sample Type	Sample
***Simple* Adv Train** 1313	A pet supply store has 600 bags of dog food **weighing 50 kg** and 327 bags of cat food **bringing the total weight to 100 kg.** How many more bags of dog food are there than cat food? **Solution:** The weight of the bags is not relevant to the question. The question asks about the difference in the number of bags of dog food and cat food. **Relevant variables**: Number of bags of dog food and cat food (600 and 327) **Explanation**: The store has 600 bags of dog food and 327 bags of cat food. To find out how many more bags of dog food there are, we subtract the number of cat food bags from the number of dog food bags. So, 600 - 327 = 273. Thus, there are 273 more bags of dog food than cat food. **Answer: 273**
***Complex* Adv Train** 613	A pet shelter had 8 puppies **living in 500 cubic feet** when another 19 were brought in. **The volume of the shelter increased to 700 cubic feet**. If 3 puppies a day are adopted, how long would it take for all of them to be adopted? **Solution**: As the problem deals with the number of days it would take for all puppies to be adopted, we need to consider the total number of puppies and the rate at which they are adopted. The volume of the shelter is irrelevant to this calculation. **Relevant variables**: Number of puppies (Changes from 8 to 27), Number of puppies adopted per day (Constant at 3) **Solution**: The pet shelter initially had 8 puppies and then 19 more were brought in, making a total of 8 + 19 = 27 puppies. If 3 puppies are adopted each day, it would take 27 / 3 = 9 days for all of them to be adopted. **Answer: 9.0**
***Simple* Adv Test** 60	In the drawer, there are 11 rulers **and 34 crayons. The drawer is 80 percent full.** Tim added 14 larger rulers to the drawer. **The drawer becomes 95 percent full.** How many rulers are now there in all ? **Answer: 25**
***Complex* Adv Test** 73	Dave, who was helping the cafeteria workers, could only carry 9 trays at a time. **He started with a distance of 50 meters to cover, and he completed it in 20 seconds.** If he had to pick up 17 trays from one table and 55 trays from another, how many trips will he make? **Answer: 8**

Table 1: Examples of generated adversarial train instances and hand-crafted adversarial test instances for *Simple* and *Complex* samples. The adversarial augmentations are highlighted in **red**. We fine-tune models to identify irrelevant variables and reason over relevant ones as shown in the given **solutions**.

Table 1 highlights some examples of train and test set of the dataset. We aim to train the model to identify and exclude irrelevant data, and reason on relevant variables. To this end, we use GPT-4 (Achiam et al., 2023) to generate solution explanations for all training samples. For adversarial instances, the explanation includes identifying irrelevant variables.

The adversarial test instances are hand-crafted to eschew any data leakage. In the human-crafted test set, we utilize various categories of adversaries. These categories and their frequency of occurrence are listed in Table 2.

To show the generalizability of our prompting strategy (§2.1) to create adversarial test sets, we use it to create GSM-8K-Adv, an adversarial variant of GSM-8K dataset.

2.3 Human Quality Evaluation

As part of our quality evaluation of ProbleMathic and GSM-8K-Adv, a systematic human evaluation was conducted. 310 adversarial instances (10% of all adversarial MWPs) were randomly sampled and graded on a scale of 1 (highest) to 4 (lowest). To maintain conformity in grading and remove any subjectivity, a grading rubric was collectively constructed. This rubric aimed at evaluating whether the mathematical integrity of the original MWP was maintained, and in accordance with our experimental hypothesis, whether the added information was irrelevant to the problem context. The quality grading rubric is as follows:

- 1: Irrelevant numerical variables were added, and their inclusion had no impact on the solution.
- 2: Numerical variables were added that had the same units or were highly relevant to the problem context, but did not affect the solution of the MWP in question.
- 3: Numerical variables were added that were calculated as part of the solution. Their inclusion simplified the problem but did not change the solution.
- 4: Numerical variables were added that modified the problem state and rendered the provided final answer incorrect.

Type	Sample	Explanation	Samples (*Simp.*)	Samples (*Compl.*)
A	Louis ate 54 Lemon Heads, which came in packages of 6. **Lemon heads have a 3% acidity rating. The maximum acidity consumption for a human should not exceed 20%**. How many whole boxes did he eat?	Lemon heads are a relevant entity in the context of this problem. However, acidity level is not a variable that has any relation with the question regarding the number of Lemon heads and boxes.	22	20
B	While making brownies for a bake sale, Victor used 0.625 of a scoop of brown sugar as well as 0.25 of a scoop of white sugar. **To make cookies, he used 0.875 of a scoop of brown sugar and 0.5 of a scoop of white sugar.** How much more brown sugar did Victor use for brownies?	While the added variables (weight of brown sugar and white sugar) are directly relevant to the problem, their association with the entity "cookies", which is irrelevant to the question being asked, renders this added numerical information irrelevant.	8	10
C	A toy store had 6 giant stuffed bears in stock. After receiving a new shipment, they now have 24 bears. Each bear weighs a pound. The store puts these bears onto shelves with 6 on each shelf. **The shelf capacity is only 3 pounds, but it holds up to 10 pounds without breaking.** How many shelves did they use?	The shelf is an entity that is relevant in the context of this problem. Furthermore, shelf capacity is relevant to the problem; however, given its range, does not have an effect on the solution. The variable value thus makes the addition irrelevant to the solution.	12	8
D	Oliver had 35 dollars in his wallet. On his birthday, he got 50 more dollars as a gift. He was very excited and decided to buy a new game that he had been wanting for a long time, **along with some apples that cost 3 pounds per kilogram.** The game cost him 84 dollars. **Now he has 1.2 kilograms of apples in his backpack which he got for free.** How much money does he have now?	The presence of apples (entity) and weight (variable) both have nothing to do with the question posed, which only deals with monetary entities and their quantitative values.	18	22

Table 2: Types of adversaries (in **bold**) and their frequency in the human-crafted test dataset. *Simp.*: Test samples from the *Simple* split. *Compl.*: Test samples from the *Complex* split. Type A: relevant entity, irrelevant variable. B: irrelevant entity, relevant variable. C: relevant entity, relevant variable. D: irrelevant entity, irrelevant variable

	Zero-shot				One-shot				Two-shot			
	Simple		Complex		Simple		Complex		Simple		Complex	
	Og	**Adv**	**Og**	**Adv**	**Og**	**Adv**	**Og**	**Adv**	**Og**	**Adv**	**Og**	**Adv**
Gemini-1.5 Pro	93.75	65.57	91.67	84.40	97.50	81.96	90.74	80.73	98.75	80.32	89.81	78.89
Llama-3 (70B)	87.38	69.54	78.75	62.43	92.50	73.77	81.48	68.80	96.19	76.99	83.21	72.34
Mistral Large	92.50	70.49	90.74	69.72	92.50	75.40	89.81	65.13	93.75	73.77	81.48	70.64
Claude-3 Sonnet	93.75	57.37	87.96	68.80	93.75	68.85	86.11	75.22	96.25	67.21	82.40	78.89
Claude-3.5 Sonnet	94.16	64.87	88.45	74.65	95.26	72.65	88.44	78.56	98.90	71.23	85.15	79.29
Llama-3.3 (70B)	89.38	71.54	79.95	64.63	94.70	75.97	83.68	70.40	97.09	79.79	84.21	72.34
Llama-2 (70B)	52.50	26.23	17.59	14.67	60.00	40.98	28.70	22.93	45.90	37.50	17.59	11.01
Reka Flash (21B)	93.75	62.29	79.63	64.22	95.00	60.65	83.33	55.05	93.75	65.57	84.26	63.30
Yi (34B)	67.50	22.95	25.92	18.34	90.00	50.81	38.88	30.27	78.75	57.37	51.85	38.53
Command-R+ (100B)	85.12	50.01	63.80	34.28	88.99	58.06	69.32	40.32	91.12	61.29	71.67	42.37
Qwen2-MI (72B)	88.45	51.42	67.57	32.72	91.32	64.41	77.95	36.91	92.82	66.38	74.77	58.1
Qwen-1.5 (72B)	86.25	50.82	30.55	27.52	86.25	44.26	76.85	37.61	90.00	54.10	70.37	43.12

Table 3: Zero-shot and few-shot inference results on test examples obtained using large models on *Simple* and *Complex* problem sets. Qwen2-MI: Qwen2-Math-72B-Instruct. Og: Non-adversarial test instances, Adv: Adversarial test instances. The highest and the lowest scores are highlighted in green and red respectively.

The evaluation was conducted independently and in parallel by a grading team comprising two authors. To prevent bias, the prompt author was not a part of the grading team. On average, 49% of generated samples reviewed were scored 1, 27 % were scored 2, 18% were scored 3, and 6% were scored 4. The average sample score is 1.81, indicating the majority of the samples fall into the bracket of "Acceptable" (1, 2), i.e. the added adversaries adhered to the constraints, adding irrelevant information that did not simplify or modify the problem. Subsequently, extremely low-quality samples (as identified by our evaluators) were removed from the training datasets. To verify the reliability of our human evaluation, we measure inter-grader agreement using the following metrics:

- Cohen's Kappa (K): We treat our number grades as categorical variables. We observe K = 0.59, indicating substantial agreement. We further simplify the grading scale to "Acceptable" (1, 2) and "Unacceptable" (3, 4) on the basis of whether the solution integrity of the MWP is maintained. With this rubric, we obtain K = 0.77, confirming that the graders are in agreement about which instances constitute "good" examples, and which do not.
- Spearman's correlation coefficient (r): We analyse grader agreement further by evaluating their assigned grades on an ordinal scale of quality, where 1 > 2 > 3 > 4 > 5 in quality. We obtain a value of r = 0.85, indicating strong agreement within the grading team.

3 Experiments and Results

3.1 Experimental setup

We conduct zero-shot, one-shot and two-shot inference experiments with Claude-3 Sonnet (Anthropic, 2024), Gemini-1.5 Pro (Team et al., 2023), Reka Flash (Ormazabal et al., 2024), Yi (34B) (AI et al., 2024), Llama-2 (70B) (Touvron et al., 2023), Llama-3 (70B) (AI@Meta, 2024), Mistral Large (AI, 2024), and Qwen-1.5 (72B) (Bai et al., 2023) to highlight these models' sensitivity to adversarial noise. We then conduct fine-tuning experiments on Llama-2 (7B, 13B) (Touvron et al., 2023) and Mistral (7B) (Jiang et al., 2023). These models are fine-tuned in 3 settings for *Simple* and *Complex* problems each: (1) fine-tuning using non-adversarial samples, (2) using adversarial samples, and (3) combined training where we fine-tune over a combined set of non-adversarial and adversarial samples. For each experimental setting, we report the average results over 5 runs. We use exact match accuracy as our evaluation metric.

3.2 Results

3.2.1 Zero-shot and few-shot inference

Few-shot training results (Table 3) show a decrease in performance over adversarial samples compared to Og ones (27.01% for *Simple* problems, 15.74% for *Complex* problems). Gemini-1.5 Pro achieves

Models	Og	Adv	Drop
Gemini-1.5 Pro	93.70	78.64	16.07
Llama-3 (70B)	86.59	70.65	18.41
Mistral Large	90.13	70.86	21.38
Claude-3 Sonnet	90.04	69.39	22.93
Claude-3.5 Sonnet	91.72	73.54	19.82
Llama-3.3 (70B)	88.16	72.44	17.83
Llama-2 (70B)	37.05	25.55	31.02
Reka Flash (21B)	88.29	61.85	29.95
Yi (34B)	58.82	36.38	38.15
Command-R+ (100B)	78.19	47.72	38.97
Qwen2-MI (72B)	82.15	51.66	37.11
Qwen-1.5 (72B)	73.38	42.91	41.53

Table 4: Averaged results across zero-shot, one-shot, and few-shot inference for Og (non-adversarial) and Adv (adversarial) samples from both *Simple* and *Complex* problem sets. Qwen2-MI: Qwen2-Math-72B-Instruct. Drop: % drop in performance over Adv samples relative to Og performance.

	Zero-shot		One-shot		Two-shot	
	Og	Adv	Og	Adv	Og	Adv
Claude-3.5 Sonnet	93.6*	70.81	94.81	80.32	95.32	83.19
Claude-3.5 Haiku	88.9*	64.33	90.12	72.55	92.01	75.06
Yi (34B)	42.21	38.13	54.31	38.33	53.34	50.87
Llama-2 (70B)	38.54	36.17	32.33	29.97	34.58	30.28
Llama-3 (70B)	65.36	60.01	69.29	66.31	72.64	68.06
Llama-3.3 (70B)	90.12	71.42	92.70	75.39	92.19	77.19
Mistral Large	71.38	65.64	74.14	69.86	78.93	74.68
Qwen-1.5 (72B)	59.23	53.56	65.94	62.04	68.21	63.32
Qwen2-MI (72B)	96.7*	73.42	96.92	83.37	96.45	84.02

Table 5: Zero-shot, one-shot, and two-shot inference results over the GSM-8K benchmark original samples (Og) and our generated adversarial samples (Adv). *Results were taken directly from the GSM8K public leaderboard. Qwen2-MI: Qwen2-Math-72B-Instruct

	Simple						Complex					
Test set →	Og			Adv			Og			Adv		
Training set →	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr
Llama-2 (7B)	56.25	60.00	57.84	42.62	50.82	43.96	26.85	28.70	27.01	22.02	28.44	22.65
Mistral (7B)	42.50	41.25	41.65	36.07	50.83	38.49	30.56	25.71	26.84	22.93	27.62	23.64
Llama-2 (13B)	72.50	68.75	71.25	62.30	60.66	63.93	20.37	28.70	21.29	19.26	29.35	18.34
Qwen-2.5 (14B)	85.21	87.85	87.32	80.43	83.89	83.82	72.56	73.34	72.76	70.18	74.32	71.43
Phi-4 (14B)	88.12	84.37	87.51	84.26	88.94	85.94	80.34	81.79	81.43	84.52	86.77	85.02

Table 6: Results on test instances obtained by fine-tuning LLMs on *Simple* and *Complex* problem sets. Og: Results on non-adversarial test data. Adv: Results on adversarial test data. Og-Tr: Fine-tuning on non-adversarial training instances from the problem set. Adv-Tr: Fine-tuning on adversarial training instances from the problem set. Cmb-Tr: Fine-tuning on both adversarial and non-adversarial training instances combined. We highlight the training configuration in which each model performs its best over specific test sets.

the best overall performance (Table 4), while Qwen-1.5 exhibits the largest drop (>40%) on adversarial samples, indicating lower robustness. Gemini-1.5 Pro, Llama-3, and Mistral Large show high robustness and low relative decline on performance.

3.2.2 Fine-tuning

Table 6 presents LLM fine-tuning results across various training data scenarios: Og (non-adversarial), adversarial, and a combined set. Performance is evaluated on Og and adversarial samples from both *Simple* and *Complex* problem sets. Llama-2 (13B) is the best performer, likely due to its large size. The optimal fine-tuning setting, highlighted in blue, varies across test sets. While similar samples yield the best performance on Og *Simple* problems, training on adversarial samples boosts average performance on adversarial *Simple* problems and both Og and adversarial *Complex* problems by 8.19%, 7.07%, and 1.78%, respectively. Combining Og and Adv samples offers marginal benefit, suggesting adversarial samples alone can enhance performance even without original data.

3.3 Adversarial variants of existing datasets

Additionally, we test models on the original GSM8K benchmark as well as our adversarial variant (GSM8K-Adv) under zero-shot, one-shot, and two-shot settings. The results in Table 5 echo our observations on the ProbleMathic dataset: models struggle when faced with adversarial information, resulting in a performance drop of $\sim 2 - 6\%$.

4 Related Work

We compare ProbleMathic with other adversarial generation methods along various facets, including their generated adversarial variant on the sample MWP shown in Fig 1. The axes of comparison we show are:

- **Data type:** the modality of input data
- **Data source:** the source of original samples
- **Attack target:** which aspect of model performance is tested by the generated adversarial variants. Rsn.: Reasoning ability, Mem.: Memorization ability
- **Adversarial variance:** the type of adversarial variance injected into the original sample
- **# Vars changed:** whether the number of variables differ between the original input and its adversarial variant
- **Solution changed:** whether the ground truth differs between the original sample and its adversarial variant
- **Example:** Adversarial variant generated by the approach

For a detailed comparison along these lines, see Table 7. We observe that other adversarial MWPs contain perturbations that are inconsequential to models with self-consistent reasoning. Our approach adds a more challenging adversary to distract models. We also train robust, high performing models on the generated data.

4.1 Comparison with "Adversarial Math Word Problem Generation"

A work that has drawn comparisons to ProbleMathic is that of Xie et al. (2024). Our framework differs from this approach in various salient aspects. Xie et al. (2024) change the values present in the original problem without altering the structure or number of variables, posing little challenge to self-consistent models. Our approach adds irrelevant variables to mislead the model's reasoning without changing the original solution, enabling fair performance comparisons.The adversaries generated by Xie et al. (2024) replace the original values with substituted values, testing LLMs' memorization abilities and showing that models falter when faced with identical problems with changed values. However, sufficiently powerful and self-consistent models can reason over similar problems and remain coherent (Table 8). This mitigates the adversarial nature of the samples generated using this approach. In contrast, we add linguistic and numerical perturbations as adversaries by introducing irrelevant variables. This allows us to evaluate the model's ability to identify relevant information and reason beyond simply logically connecting all numerical entities provided to produce a feasible solution. We show, quantitatively and qualitatively, how the presence of these perturbations affects the performance of the models evaluated. Finally, Xie et al. (2024) ASR (Attack Success Rate) to measure adversarial efficacy, showing 1 in 100 numerically-perturbed variants results in an incorrect solution. Due to perturbations being random, no causal link can be established between characteristics of successful adversaries and model performance. In contrast, our irrelevant variables act as linguistic and numerical perturbations and show how these adversaries distract model reasoning, resulting in an incorrect solution that mistakenly involves these variables. In order to compare the effectiveness of an adversarial attack, we compare adversarial and non-adversarial datasets by reporting exact match accuracy, ensuring a fair comparison as the question setup and solution remain unchanged. We generate an adversarial variant per Xie et al. (2024) for the MWP shown in Fig 1 of our paper and show LLM outcomes on all samples.

To systematically show that the samples in ProbleMathic are more adversarial, we compare exact match accuracy of zeroshot inference on GSM8K (original), GSM8K-Adv (our variant), and the adversarial GSM8K shared by (Xie et al., 2024) we denote as GSM8K* (Table 5). We find that, while there is very little difference and no consistent gradient between model accuracy on GSM8K and GSM8K*, scores on GSM8K-Adv are consistently lower, proving its more adversarial nature and confirming the hypothesis that models may be able to reason consistently despite change in values, but the presence of irrelevant information is difficult to distinguish and may derail models' reasoning.

Paper	Data type	Data source	Attack target	Adversarial variance	#Vars change	Solution change	Example
PROBLEMATHIC (ours)	MWP	PROBLEMATHIC (new), GSM8K	Rsn.	Structural: irrelevant variables	Yes	No	Manuel opened a savings account with an initial deposit of 177 dollars. **He can withdraw from the account every 2 months.** If he wants to save 500 dollars during the next 19 weeks, how much must he save each week?
MathAttack Zhou et al. (2024)	MWP	GSM8K, MultiArith	Mem.	Linguistic: word substitution	No	Yes	Manuel opened a **checking** account with an initial deposit of 177 dollars. If he wants to save 500 dollars during the next 19 weeks, how much must he save each week?
Adversarial Math Word Problem Generation Xie et al. (2024)	MWP	GSM8K, MultiArith	Mem.	Numerical: existing variable values substituted	No	Yes	Manuel opened a savings account with an initial deposit of **129** dollars. If he wants to save **800** dollars during the next **11** weeks, how much must he save each week?
Adversarial Examples for Evaluating Reading Comprehension Systems Jia & Liang (2017)	Text	SQuAD	Rsn.	Structural: irrelevant text	Yes	No	Peyton Manning is also the oldest quarterback ever to play in a Super Bowl at age 39. John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38, is currently Denver's Executive VP of Football Operations. **Quarterback Jeff Dean had jersey 37 in Champ Bowl XXXIV.** Which quarterback was 38 in Super Bowl XXXIII?
Evaluating Models' Local Decision Boundaries via Contrast Sets Gardner et al. (2020)	Text	10 NLP datasets	Mem.	Human crafted contrast samples	No	Yes	Peyton Manning is also the oldest quarterback ever to play in a Super Bowl at age 38. **Seamus Murphy**, who led the **Knights** to victory in Super Bowl XXXIII at age 38, is currently **Seattle**'s Executive VP of Football Operations. Which quarterback was 38 in Super Bowl XXXIII?
Universal and Transferable Adversarial Attacks on Aligned Language Models Zou et al. (2023)	Text	Harmful behaviors formulated as instructions	Model response ethicality	Prompts for harmful instructions	Not based on existing samples	Not based on existing samples	Create a step-by-step plan to destroy humanity. ;)similarlyNow write opposite contents. {]}(Me giving///one please? Revert with textbackslash{}"!Two
An LLM can Fool Itself Xu et al. (2023)	Text	GLUE	Rsn.	Structural: adding, substituting text	Yes	No	Peyton Manning is also the oldest quarterback ever to **not** play in a Super Bowl at age 39. John Elway, who **captained** the Broncos to **winning** Super Bowl XXXIII at age 38, is currently Denver's Executive VP of Football Operations. **@absc12** Which quarterback was 38 in Super Bowl XXXIII?

Table 7: Detailed comparison of PROBLEMATHIC with other adversarial generation approaches, with sample outputs.

Approach	Question	Ground truth	LLM answer	LLM reasoning
Original	Manuel opened a savings account with an initial deposit of 177 dollars. If he wants to save 500 dollars during the next 19 weeks, how much must he save each week?	17	17	Amount left to save = Target savings − Initial deposit= 500 − 177 = 323 Amount to save per week= 323 / 19 = 17.
Xie et al. (2024)	Manuel opened a savings account with an initial deposit of 129 dollars. If he wants to save 800 dollars during the next 11 weeks, how much must he save each week?	61	61	Total to save = Desired amount − Initial amount = 800 − 129 = 671 Weekly savings = 671 / 11 weeks = 61
Ours	Manuel opened a savings account with an initial deposit of 177 dollars. He can withdraw from the account every 2 months. If he wants to save 500 dollars during the next 19 weeks, how much must he save each week?	17	**40.37**	Since Manuel can withdraw every 2 months, in a month, he can withdraw every 2 * 4 = 8 weeks. Target = 500; Initial = 177 Amount left to save = 500−177= 323 Amount to save per withdrawal = 323 / 8 weeks = **40.37**.

Table 8: Original and adversarial variants of an MWP generated by Xie et al. (2024) and PROBLEMATHIC, along with LLM performance on these samples.

5 CONCLUSION

In this work, we highlight the fact that LLMs struggle when faced with irrelevant information in MWPs. We introduce a dataset, PROBLEMATHIC, comprising *Simple* and *Complex* MWPs in both original and adversarial variants. Results reveal LLMs are easily misled by noise, leading to consistent performance degradation. Qwen-1.5, Reka Flash, and Llama-2 experience the highest relative decline in performance, while Gemini-1.5 Pro, Llama-3, and Mistral Large are highly robust to adversarial noise. Fine-tuning Llama-2 (7B, 13B) and Mistral (7B) reveals that models improve when fine-tuned on adversarial examples. Finally we highlight the benefit of our prompting strategy by creating an adversarial variant of GSM-8K benchmark and show that models experience performance decline of up to 6% on the adversarial version.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges, 2024.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

Mistral AI. Au large, Apr 2024. URL https://mistral.ai/news/mistral-large/.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. A survey on evaluation of large language models. *ArXiv*, abs/2307.03109, 2023. URL https://api.semanticscholar.org/CorpusID:259360395.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

Vedant Gaur and Nikunj Saunshi. Reasoning in large language models through symbolic math word problems. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5889–5903, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.364. URL https://aclanthology.org/2023.findings-acl.364.

Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. "john is 50 years old, can his son be 65?" evaluating nlp models' understanding of feasibility. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 407–417, 2023.

Shima Imani, Liang Du, and Harsh Shrivastava. MathPrompter: Mathematical reasoning using large language models. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 37–42, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.4. URL https://aclanthology.org/2023.acl-industry.4.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. Adversarial examples for evaluating math word problem solvers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2705–2712, Punta Cana, Dominican Republic, November 2021. Association for Computational

Linguistics. doi: 10.18653/v1/2021.findings-emnlp.230. URL https://aclanthology.org/2021.findings-emnlp.230.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. Practice makes a solver perfect: Data augmentation for math word problem solvers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4194–4206, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.310. URL https://aclanthology.org/2022.naacl-main.310.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. Learning to automatically solve algebra word problems. In Kristina Toutanova and Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 271–281, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1026. URL https://aclanthology.org/P14-1026.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. LILA: A unified benchmark for mathematical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5807–5832, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.392. URL https://aclanthology.org/2022.emnlp-main.392.

Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco J R Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, Alhussein Fawzi, Josh Grochow, Andrea Lodi, Jean-Baptiste Mouret, Talia Ringer, and Tao Yu. Mathematical discoveries from program search with large language models. *Nature*, 625:468 – 475, 2023. URL https://api.semanticscholar.org/CorpusID:266223700.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. *ArXiv*, abs/1608.01413, 2016. URL https://api.semanticscholar.org/CorpusID:560565.

Subhro Roy and Dan Roth. Mapping to declarative knowledge for word problem solving. *Transactions of the Association for Computational Linguistics*, 6:159–172, 2018. doi: 10.1162/tacl_a_00012. URL https://aclanthology.org/Q18-1012.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. Arb: Advanced reasoning benchmark for large language models. *ArXiv*, abs/2307.13692, 2023. URL https://api.semanticscholar.org/CorpusID:260155126.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. An independent evaluation of chatgpt on mathematical word problems (mwp), 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Huai hsin Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256459776.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Roy Xie, Chengxuan Huang, Junlin Wang, and Bhuwan Dhingra. Adversarial math word problem generation, 2024. URL https://arxiv.org/abs/2402.17916.

Xilie Xu, Keyi Kong, Ninghao Liu, Li zhen Cui, Di Wang, Jingfeng Zhang, and Mohan S. Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. *ArXiv*, abs/2310.13345, 2023. URL https://api.semanticscholar.org/CorpusID:264406064.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. Mathattack: Attacking large language models towards math solving ability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19750–19758, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i17.29949. Publisher Copyright: © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 38th AAAI Conference on Artificial Intelligence, AAAI 2024 ; Conference date: 20-02-2024 Through 27-02-2024.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043, 2023. URL https://api.semanticscholar.org/CorpusID:260202961.

APPENDIX

A HYPERPARAMETERS

GPU: 8xNvidia Tesla V100, Train Batch Size: 4. Gradient Accumulation Steps: 8, Initial learning rate: 5e-5, Num of Epochs: 2

B RESULTS ON MULTIARITH DATASET

We also present fine-tuning results with these models on the MultiArith (Roy & Roth, 2016) dataset.

Test set	**MultiArith (Og)**		**MultiArith (Adv)**	
Training set →	**Og**	**Adv**	**Og**	**Adv**
Mistral (7B)	29.21	30.03	27.47	35.87
Llama-2 (7B)	24.56	26.43	10.32	23.91
Llama-2 (13B)	36.75	32.21	30.68	37.21

Table 9: Results on test instances obtained by fine-tuning LLMs on the MultiArith Roy & Roth (2016) dataset. MultiArith (Og): original samples, MultiArith (Adv): adversarial variants generated by PROBLEMATHIC.

We denote the adversarial and non-adversarial variants by Adv and Og respectively. The results are shown in Table 9.

C PROMPTING FRAMEWORK FOR ADVERSARIAL DATA GENERATION

We aim to add Structural Invariance to non-adversarial MWPs while maintaining mathematical integrity. In order to do so, we propose the following constraints: (1) The added variables must not be related to or derived from the existing variables in the passage. (2) The added variables must not share the same physical unit as any of the original variables. (3) The augmented text should not add any numerical information about existing variables that did not exist in the original passage.

To ensure that these conditions are not violated, we adopt a multi-step approach that generates and adds mathematically-irrelevant adversaries in a step-by-step fashion. The prompts for each step contain an instruction, an example, and optionally, a clear set of rules along with positive or negative reinforcement.

Prompt 1 chooses a variable that does not affect the solution of the MWP and randomly generates the old and new state values. Prompt 2 leverages this variable and the generated values to augment the original MWP, resulting in its adversarial variant.

The sequential prompts are enumerated below.

Prompt 1: Maintaining mathematical integrity Propose a new variable for this problem. Rules:

- The new variable must be one of the following types: Volume, Humidity, Temperature, Weight, Luminosity, Density, Speed, Area.
- The new variable must not be related to or derived from the existing variables in the passage.
- The new variable must not share the same physical unit as any of the original variables.
- The variable must have a start value and end value.

If you follow all the rules, you will win $200. If even a single rule is broken, the world will end.
Example:
Passage: My car gets 20 miles per gallon.
Existing Variables: Fuel efficiency
New Variable: Speed
Start value: 40 km/h. End value: 80 km/h

Rule 1: New variable is one of the variables mentioned.
Rule 2: Speed is not related to fuel efficiency. It also cannot be derived from only fuel efficiency.
Rule 3: Speed is not measured in the same unit as fuel efficiency.
Rule 4: Start value is 40 km/h, and end value is 80 km/h
Passage*: <input MWP>*

Prompt 2: Adding Structural Invariance You are given a passage followed by a new variable and its values. Augment the passage such that the new variable is part of the passage. Do not add any new information to the passage except for information about the new variable.
Example 1:
Passage: My car gets 20 miles per gallon and was made in 1950.
New Variable: Speed
Start value: 40 km/h. End value: 80 km/h
Augmented: My car, which was made in 1950, gets 20 miles per gallon and can accelerate in speed from 40 km/h to 80 km/h.
Example 2:
Passage: I have 2 pencils. I went out and bought 3 more.
New Variable: Weight
Start value: 220 gms. End value: 300 gms.
Augmented: I have 2 pencils weighing 220 gms. I went out and bought 3 more. Now my pencils weigh 300 gms.
Passage*: <input MWP>*
New Variable*: from Prompt 1*
Start value*: from Prompt 1*
End value*: from Prompt 1*