

## 718 Appendix

719 The dataset and code to run all experiments are provided in this [repository](#).

### 720 8 MWP-MISTAKE Dataset

721 MWP-MISTAKE dataset is curated using 4 different types of well-known datasets. Below are the details  
722 of each of the datasets.

- 723 • GSM-8K [10]: GSM-8K is a dataset of diverse grade school math word problems created by  
724 human writers, involving basic arithmetic operations. Released in November 2021.
- 725 • MATH [16]: The MATH dataset is divided into seven categories, each with five difficulty  
726 levels. For our study, we used levels 1, 2, and 3 from the algebra and counting and probability  
727 categories. Released in November 2021.
- 728 • MATHBENCH [20]: MATHBENCH is a recent dataset with questions divided by educa-  
729 tional stages, from basic arithmetic to college levels. For our experiment, we chose middle  
730 and high-school-level single-choice multiple-choice questions. Released in May 2024.
- 731 • JEEBENCH [6]: JEEBENCH is a challenging benchmark dataset for evaluating LLM  
732 problem-solving abilities, containing 515 pre-engineering math, physics, and chemistry  
733 problems from the IIT JEE-Advanced Exam. For our experiment, we chose mathematics  
734 single-choice questions. Released in October 2023.

#### 735 8.1 Prompts to curate reasoning steps in MWP-MISTAKE dataset

736 GSM-8K and MATH already contain MWP questions, a chain of thought reasoning steps  
737 and a final answer. To curate chain of thought reasoning step for MATHBENCH and  
738 JEEBENCH we made use of GPT-4. While prompting GPT-4 we made sure that rea-  
739 soning steps did not contain the final answer, so that final answer is not picked di-  
740 rectly from the reasoning step. Listing 1 prompt is used to curate the reasoning steps.

```
1 Strictly follow the below conditions.  
2 1. Output format: \nReasoning Chain: \nFinal Answer:  
3 2. Reasoning Chain should be separated by a new line only.  
4 3. Reasoning chain cannot have the final answer. (Replace the  
741 final answer in the reasoning chain with its calculation  
or ####)  
5 4. Do not include any additional information in the final  
answer (only the answer).
```

Listing 1: Prompt to curate reasoning chain without answers.

742 Table 5 shows examples of default reasoning steps from GSM-8K dataset.

Table 5: Example of rule based incorrect reasoning step (GSM-8K dataset)

Question	Gerald spends \$100 a month on baseball supplies. His season is 4 months long. He wants to use the months he's not playing baseball to save up by raking, shoveling, and mowing lawns. He charges \$10 for each. How many chores does he need to average a month to save up for his supplies?
Final Answer	5
Gold Reasoning step	He needs to save up \$400 because $4 \times 100 = 400$ He has 8 months to earn this money because $12 - 4 = 8$ He needs to earn \$50 a month because $400 / 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$
Shuffle reasoning step	He needs to earn \$50 a month because $400 / 8 = 50$ He needs to save up \$400 because $4 \times 100 = 400$ He needs to do 5 tasks a month because $50 / 10 = 5$ He has 8 months to earn this money because $12 - 4 = 8$
Delete reasoning step	He needs to save up \$400 because $4 \times 100 = 400$ He needs to earn \$50 a month because $400 / 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$
Shuffle numerical values	He needs to save up \$400 because $4 \times 100 = 400$ He has 50 months to earn this money because $8 - 8 = 4$ He needs to earn \$12 a month because $400 / 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$
Replace numerical values	He needs to save up \$400 because $4 \times 100 = 400$ He has 8 months to earn this money because $12 - 4 = 8$ He needs to earn \$6 a month because $32 / 8 = 50$ He needs to do 76 tasks a month because $50 / 10 = 5$
Shuffle Operations	He needs to save up \$400 because $4 \times 100 = 400$ He has 8 months to earn this money because $12 * 4 = 8$ He needs to earn \$50 a month because $400 - 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$
Insert Random Reasoning step	He needs to save up \$400 because $4 \times 100 = 400$ <b>Therefore, Faye has \$60 - \$30 = \$30 left.</b> He has 8 months to earn this money because $12 - 4 = 8$ He needs to earn \$50 a month because $400 / 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$

## 743 9 SLMs reasoning steps

744 SLMs were used to generate chain of thought (COT) reasoning step and final answers for all dataset  
745 questions. Each model Llama-2-7b-chat, Mixtral-8x7B, Phi-3-mini were prompted using Listing 1  
746 to curate the reasoning step without an answer. If the final answer was incorrect we filtered out the  
747 reasoning steps as incorrect.

748 Table 6 shows examples of SLM incorrect reasoning steps from GSM-8K dataset.

## 749 10 Task T1 and T2

750 Task T1 evaluates the model's ability to detect mistakes rectify them and derive the correct answer.  
751 Listing 2 was used in a few shot settings for task T1.

752  
753 1 You are a mathematics educator with a deep understanding of  
754 elementary and middle school mathematics. You are  
755 experienced in teaching multi-step problem-solving  
756 techniques and have a knack for breaking down complex  
757 problems into manageable steps. Your expertise lies in  
758 basic arithmetic operations such as addition, subtraction,  
759 multiplication, and division. You can provide clear, step-  
760 by-step solutions to mathematical problems that require  
761 multi-step reasoning.  
762 2

Table 6: Example of incorrect reasoning steps generated using SLM’s (GSM-8K dataset)

Question	Gerald spends \$100 a month on baseball supplies. His season is 4 months long. He wants to use the months he’s not playing baseball to save up by raking, shoveling, and mowing lawns. He charges \$10 for each. How many chores does he need to average a month to save up for his supplies?
Final Answer	5
Gold Reasoning step	He needs to save up \$400 because $4 \times 100 = 400$ . He has 8 months to earn this money because $12 - 4 = 8$ He needs to earn \$50 a month because $400 / 8 = 50$ He needs to do 5 tasks a month because $50 / 10 = 5$ #### 5
Llama-2-7b-chat Incorrect reasoning step	Gerald spends \$100 on baseball supplies per month. Season is 4 months long. To save up for baseball supplies, Gerald needs to earn more than the cost of supplies per month. Gerald charges \$10 for each chore he does. To calculate the number of chores Gerald needs to do per month, divide the cost of supplies by the amount he charges per chore.
Llama-2-7b-chat Predicted Answer	""
Mixtral-8x7B Incorrect reasoning step	Gerald spends \$100 a month for 4 months, so his total spending for the season is $100 \times 4 = \$400$ Now, we need to find out how many chores he needs to do to earn this amount He charges \$10 per chore, so we divide the total amount by the cost per chore: $400 / 10 = ####$
Mixtral-8x7B Predicted Answer	40
Phi-3-mini Incorrect reasoning step	Gerald’s total expenditure for baseball supplies = \$100 per month * 4 months = \$400 Gerald earns \$10 per chore. To find out how many chores he needs to do to save up \$400, we divide his total expenditure by the amount he earns per chore. = $\$400 / \$10 = ####$
Phi-3-mini Predicted Answer	40

```

763 3 You are provided with a mathematical question and a step-by-
764     step solution along with it. The solution might have some
765     mistakes. Identify if the solution is correct or incorrect.
766     If the solution is correct, output the final answer with
767     the help of the solution provided. If the solution is
768     incorrect, correct the existing solution and determine the
769     final answer with the help of the corrected solution.
770 4 Reasoning chain Correct (Yes/No):
771 5 Corrected reasoning chain or NA:
772 6 Final answer (just the number):

```

Listing 2: Prompt for Task T1

774 Task T2 evaluates the model’s ability to detect mistake and solve MWP based on the provided  
775 reasoning step. Listing 3 was used in a few shot setting for task T2. Here we insure that final answer  
776 is generated with the help of the reasoning steps provided, which may or may not be correct.

```

777
778 1 You are a mathematics educator with a deep understanding of
779     elementary and middle school mathematics. You are
780     experienced in teaching multi-step problem-solving
781     techniques and have a knack for breaking down complex
782     problems into manageable steps. Your expertise lies in
783     basic arithmetic operations such as addition, subtraction,
784     multiplication, and division. You can provide clear, step-
785     by-step solutions to mathematical problems that require
786     multi-step reasoning.
787 2
788 3 You are provided with a mathematical question and a step-by-
789     step solution along with it. The solution might have some
790     mistakes. Identify if the solution is correct or incorrect
791     and output the final answer based on the provided solution.
792 4 Reasoning chain Correct (Yes/No):
793 5 Final answer (just the number):
794

```

Listing 3: Prompt for Task T2

795 **11 T2 Results**

796 Task T2 evaluates the performance in deriving the final answer based on the reasoning step which  
 797 may or may not be correct. In task T2 we do not instruct the model to correct the reasoning step, and  
 798 calculate the final answer based on the provided reasoning step. Due to which we see a significant drop  
 799 in performance between Task T1 and Task T2. Table 7 presents the mistake detection performance  
 800 (F1 score) of all the models with Task T2 and Table 8 presents the performance in deriving the final  
 answer (F1 Score) of all the models.

Table 7: Mistake Detection Performance (F1 score) on MWP-MISTAKE dataset for Task T2. (D-Default reasoning steps, SM-Smaller model reasoning steps) (Bold: Best, Underline:Second best)

Model	GSM-8K		MATH		MATHBENCH		JEEBENCH		Average		
	D	SM	D	SM	D	SM	D	SM	D	SM	Overall
GPT-4o	---	---	---	---	---	---	---	---	---	---	---
GPT-4	0.67	0.61	0.75	0.76	0.48	0.88	0.76	0.85	0.66	0.78	0.72
GPT-3.5Turbo	0.58	0.40	0.69	0.42	0.33	0.24	0.51	0.41	0.53	0.36	0.45
Llama-2-7b-chat	0.11	NA	0.22	NA	0.11	NA	0.75	NA	0.30	NA	0.30
Mixtral-8x7B	0.69	NA	0.75	NA	0.60	NA	0.76	NA	0.70	NA	0.70
Phi-3-mini	0.56	NA	0.52	NA	0.46	NA	0.54	NA	0.52	NA	0.52
Claude-3-Opus	---	---	---	---	---	---	---	---	---	---	---

801

Table 8: Performance in deriving correct answers (F1 score) on MWP-MISTAKE dataset for Task T2. (D-Default reasoning steps, SM-Smaller model reasoning steps) (Bold: Best, Underline:Second best)

Model	GSM-8K		MATH		MATHBENCH		JEEBENCH		Average		
	D	SM	D	SM	D	SM	D	SM	D	SM	Overall
GPT-4o	---	---	---	---	---	---	---	---	---	---	---
GPT-4	0.99	0.65	0.72	0.48	0.82	0.27	0.39	0.29	0.73	0.42	0.57
GPT-3.5Turbo	0.85	0.26	0.66	0.31	0.67	0.16	0.48	0.20	0.67	0.23	0.45
Llama-2-7b-chat	0.84	NA	0.33	NA	0.44	NA	0.36	NA	0.49	NA	0.49
Mixtral-8x7B	0.91	NA	0.64	NA	0.68	NA	0.11	NA	0.58	NA	0.58
Phi-3-mini	0.92	NA	0.62	NA	0.65	NA	0.49	NA	0.67	NA	0.67
Claude-3-Opus	---	---	---	---	---	---	---	---	---	---	---

802 **12 Model Used**

803 Below are brief details of the models we have used for benchmarking our MWP-MISTAKE dataset.

- 804 1. **GPT-4o:** GPT-4o is a multimodal model by OpenAI, and it has the same high intelligence  
 805 as GPT-4 Turbo but is much more efficient—it generates text 2x faster and is 50% cheaper.  
 806 Additionally, GPT-4o has the best vision and performance across non-English languages of  
 807 any OpenAI model. Last training data: October 2023.
- 808 2. **GPT-4:** GPT-4 is a large multimodal model by OpenAI that can solve difficult problems  
 809 with greater accuracy than any of OpenAI previous models, thanks to its broader general  
 810 knowledge and advanced reasoning capabilities. Last training data: September 2021.
- 811 3. **GPT-3.5Turbo:** GPT-3.5Turbo is a large language model by OpenAI GPT-3.5 that can  
 812 understand and generate natural language or code and has been optimized for chat using  
 813 the Chat Completions API but work well for non-chat tasks as well. Last training date:  
 814 September 2021.
- 815 4. **Claude-3-Opus:** Claude-3-Opus is Anthropic’s most capable and intelligent model yet,  
 816 ideal for navigating complex tasks like in-depth analysis, research, and task automation.  
 817 Last training data: August 2023.
- 818 5. **Llama-2-7b-chat:** Llama 2 is a collection of pretrained and fine-tuned generative text  
 819 models ranging in scale from 7 billion to 70 billion parameters from meta. This is the 7B  
 820 fine-tuned model, optimized for dialogue use cases. Training date: September 2022.
- 821 6. **Mixtral-8x7B:** Mixtral is a Mixture of Experts (MoE) model with 8 experts per MLP,  
 822 with a total of 45 billion parameters. Despite the model having 45 billion parameters, the  
 823 compute required for a single forward pass is the same as that of a 14 billion parameter  
 824 model. This is because even though each of the experts have to be loaded in RAM (70B

825 like ram requirement) each token from the hidden states are dispatched twice (top 2 routing)  
 826 and thus the compute (the operation required at each forward computation) is just 2 X  
 827 sequence\_length.

828 7. **Phi-3-mini:** The Phi-3-Mini-128K-Instruct is a 3.8 billion-parameter by microsoft,  
 829 lightweight, state-of-the-art open model trained using the Phi-3 datasets. This dataset  
 830 includes both synthetic data and filtered publicly available website data, with an emphasis  
 831 on high-quality and reasoning-dense properties. Last training data: October 2023.

### 832 13 METEOR and BertScore results

833 BertScore computes a similarity score for each token in the candidate sentence with each token  
 834 in the reference sentence using the BERT embeddings. Metric for Evaluation of Translation with  
 835 Explicit Ordering (METEOR) score is a metric that measures the quality of generated text based on  
 836 the alignment between the generated text and the reference text. The metric is based on the harmonic  
 837 mean of unigram precision and recall, with recall weighted higher than precision.

838 Table 9 and Table 10 present the BertScore and Meteor Score respectively for all the datasets across  
 839 all models. We observed that these two metric evaluations where not fully able to capture the nuance  
 840 capabilities of LLMs in rectifying the mistakes within reasoning steps. This can be seen in the  
 841 results. GPT-4o has a consistently high performance across all the dataset, but when you compare the  
 842 BERTScore between the corrected reasoning step and ground truth reasoning step you see the rest of  
 843 the models clearly performing better than GPT-4o. GPT-4 has performed better than GPT-3.5Turbo  
 844 in most datasets.

Table 9: BERTscores for correct and incorrect final answers derived after mistake rectification across all models and datasets.

Datasets	Models	GPT-4o		GPT-4		GPT-3.5Turbo		Llama-2-7b-chat		Mixtral-8x7B		Phi-3-mini	
		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GSM-8K	D	0.95	0.91	0.98	0.93	0.97	0.95	0.96	0.98	0.97	0.94	0.94	0.91
	SM	0.83	0.82	0.84	0.82	0.84	0.82	NA	NA	NA	NA	NA	NA
MATH	D	0.88	0.90	0.96	0.93	0.95	0.93	0.96	0.88	0.95	0.92	0.90	0.87
	SM	0.84	0.80	0.83	0.81	0.84	0.81	NA	NA	NA	NA	NA	NA
MATHBENCH	D	0.88	0.83	0.97	0.95	0.97	0.94	0.90	0.89	0.96	0.95	0.93	0.90
	SM	0.82	0.82	0.85	0.82	0.84	0.83	NA	NA	NA	NA	NA	NA
JEEBENCH	D	0.89	0.89	0.88	0.87	0.94	0.95	0.86	0.82	0.85	0.87	0.70	0.85
	SM	0.86	0.87	0.85	0.86	0.78	0.86	NA	NA	NA	NA	NA	NA

Table 10: Meteor Score for correct and incorrect final answers derived after mistake rectification across all models and datasets.

Datasets	Models	GPT-4o		GPT-4		GPT-3.5Turbo		Llama-2-7b-chat		Mixtral-8x7B		Phi-3-mini	
		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GSM-8K	D	0.81	0.54	0.92	0.62	0.88	0.77	0.87	0.83	0.85	0.74	0.77	0.66
	SM	0.33	0.27	0.37	0.31	0.37	0.32	NA	NA	NA	NA	NA	NA
MATH	D	0.48	0.54	0.76	0.70	0.76	0.67	0.78	0.59	0.73	0.66	0.55	0.48
	SM	0.32	0.28	0.30	0.26	0.33	0.28	NA	NA	NA	NA	NA	NA
MATHBENCH	D	0.55	0.35	0.82	0.63	0.82	0.68	0.49	0.57	0.81	0.68	0.67	0.53
	SM	0.33	0.30	0.32	0.25	0.32	0.29	NA	NA	NA	NA	NA	NA
JEEBENCH	D	0.37	0.31	0.30	0.22	0.49	0.54	0.15	0.13	0.53	0.46	0.20	0.25
	SM	0.28	0.26	0.21	0.21	0.08	0.25	NA	NA	NA	NA	NA	NA

### 845 14 Average reasoning Step Length

846 We noticed that the average word length of rectified reasoning for correct and incorrect for GPT-4o  
 847 was higher than other models. Table 11 presents the average word length of the rectified reasoning  
 848 step for all datasets across the models.

### 849 15 Data contamination and Memorization

850 Two different types of prompts were used to test out memorization of individual instances of the  
 851 dataset. The reasoning step was split into two halves, and models were prompted in two different

Table 11: Average length of rectified reasoning steps on MWP-MISTAKE dataset

Model	GSM-8K		MATH		MATHBENCH		JEEBENCH		Average		
	D	SM	D	SM	D	SM	D	SM	D	SM	Overall
GPT-4o	100.14	131.47	147.50	182.69	312.11	323.45	647.66	619.09	301.85	314.18	308.01
GPT-4	66.59	122.24	79.32	121.59	146.54	140.43	356.71	322.53	162.29	176.69	169.49
GPT-3.5Turbo	66.58	126.30	94.17	124.56	140.50	177.36	670.34	338.53	242.90	191.69	217.29
Llama-2-7b-chat	44.73	NA	113.35	NA	177.67	NA	137.05	NA	118.20	NA	118.20
Mixtral-8x7B	63.04	NA	88.26	NA	140.57	NA	402.79	NA	173.67	NA	173.67
Phi-3-mini	84.92	NA	115.10	NA	172.57	NA	293.90	NA	166.62	NA	166.62
Claude-3-Opus	62.18	138.91	70.60	134.05	144.85	192.84	561.88	438.44	209.88	226.06	217.97

ways (general and guided) to complete the second half of the reasoning step. To detect contamination we evaluated the overlap score between generated and reference instances using ROUGE-L Score. Listing 4 and Listing 5 shows the prompt used for guided and generation instructions respectively.

```

855 You are provided with a full question and the first answer
856 piece from the test split of the GSM8K dataset. Finish the
857 second answer piece of the instance as exactly appeared in
858 the dataset. Only rely on the original form of the instance
859 in the dataset to finish the second answer piece. Output
860 only the second answer piece.
861
862

```

Listing 4: Guided instruction for dataset GSM8K

```

863 Based on the provided question, finish the second answer piece
864 based on the first answer piece, such that these two pieces
865 become a single instance answer. Output only the second
866 answer piece.
867
868

```

Listing 5: General instruction for dataset GSM8K

Here GSM-8K and test are the extra information provided for the model to uniquely identify instances from the source dataset and complete the reasoning step.

Table 12 presents the complete result for the average ROUGE-L score of guided and general for all datasets across all models.

Table 12: Rouge L score between guided and general instructions on MWP-MISTAKE dataset

Datasets	Models	GPT-4o		GPT-4		GPT-3.5Turbo		Llama-2-7b-chat		Mixtral-8x7B		Phi-3-mini	
		Guided	General	Guided	General	Guided	General	Guided	General	Guided	General	Guided	General
GSM-8K	D	0.57	0.44	<b>0.67</b>	0.56	<b>0.53</b>	0.49	0.26	0.28	0.46	0.44	0.32	0.32
	SM	0.55	0.51	0.57	0.55	0.49	0.47	0.30	0.32	0.55	0.50	0.42	0.41
MATH	D	0.44	0.25	0.52	0.48	0.39	0.38	0.25	0.26	0.39	0.32	0.26	0.27
	SM	0.51	0.38	0.54	0.54	0.45	0.44	0.30	0.29	0.48	0.46	0.38	0.39
MATHBENCH	D	0.43	0.41	0.48	0.46	0.38	0.36	0.26	0.28	0.36	0.36	0.30	0.30
	SM	0.40	0.38	0.43	0.42	0.39	0.38	0.30	0.33	0.40	0.38	0.29	0.30
JEEBENCH	D	0.43	0.39	0.42	0.40	0.34	0.33	0.27	0.25	0.38	0.34	0.33	0.31
	SM	0.32	0.29	0.34	0.35	0.31	0.24	0.22	0.25	0.26	0.27	0.20	0.22

872

## 873 16 Running Experiment Multiple Times

While running experiments on all models (LLMs and SLMs) we used the default hyperparameters to generate tokens. To test out the model we reran the model with the same hyperparameter to test out any change in accuracy.

## 877 17 Output from each model

The raw output of each model has been provided in this [repository](#). Additional details are present in the README.md file of the repository.