CS2916 Spring 2024 Project: Mastering Efficient Reasoning with Focusing on Numbers and Hybrid

Yichao Zhong*, Yongshan Chen*[*]
521021910154, 521030910329
Shanghai Jiao Tong University

Overview: This project aims to achieve efficient reasoning (i.e., mathematical reasoning) by focusing on a carefully designed reasoning process. Specific requirements are as follows:

- Suggested experiments settings: (i) Benchmark: GSM8K; (ii) Base model: Mistral-7B.
- **Assignment 3**: (i) Reproduce the results of Mistral-7B on GSM8K (refer to the results in (Fan et al., 2024)), (ii) Perform supervised learning by removing the "Solution Step" from the training set (e.g., *Question: Natalia sold clips to 48 ... April and May? Answer: 72.*), (iii) Replace the solution steps in the training set (refer to (Pfau et al., 2024)), and (iv) List the results and provide insightful analyses.
- **Assignment 4**: Please continue to boost performance through (i) training using both "Question → Solution, Answer" and "Question → Answer" data formats combined, (ii) constructing training samples with varying solution lengths using GPT-3.5 (or other powerful models) for training, (ii) Utilizing a larger model for training, and (iv) other approaches as necessary.

Please conduct experiments, complete the above assignments, and write this report.
Note: This frame can be commented out when the report is finally submitted.

Abstract

This paper introduces an innovative methodology designed to enhance the reasoning efficiency of language models, with a specific emphasis on addressing mathematical problem-solving challenges within educational frameworks. We embarked on an investigative journey into the ramifications of leveraging varied dataset configurations—namely Vanilla, Reformatted, and Only Answer formats extracted from the General School Mathematics 8,000 (GSM8k) dataset. This exploration is aimed at understanding the influence of meticulously structured reasoning sequences and the inclusion of explicit numerical data representations on the performance outcomes of these computational models.

The cornerstone of our research methodology is the deployment of the Base Model Mistral7B-v0.1. This model serves as our experimental platform, which we have augmented through the application of custom Soft Prompt Tuning (SFT). This tuning process involves the strategic use of various structured inputs designed to optimize the model's reasoning pathways. By focusing on a dual-dataset strategy that amalgamates the Only Answer and Reformatted datasets, our approach not only facilitates a deeper understanding of data structuring effects but also introduces a novel prompting mechanism. This mechanism utilizes the capabilities of the GPT-3.5 model to generate solutions with a clear emphasis on minimizing redundancy in the answers provided.

Our preliminary findings are encouraging. They indicate that the deliberate reformatting of data to highlight step-by-step logical deductions and a numerical emphasis substantially bolsters the accuracy and reasoning acuity of the model. This enhancement is particularly notable when the model confronts complex mathematical problem-solving scenarios. The dual-dataset approach allows for a comprehensive evaluation of how different data presentations affect model performance, providing insights that are critical for refining the educational applications of language models.

Furthermore, our innovative prompting strategy, which employs GPT-3.5, proves to be effective in fostering a reduction in redundant data processing, thereby streamlining the reasoning process. This strategy not

[*] Codes are aviliable at https://github.com/z-taylcr7/FoN

only aids in producing more precise and concise answers but also exemplifies how targeted prompting can be leveraged to improve the utility of language models in educational settings.

The implications of this study are profound, suggesting that the careful structuring of data and the tailored use of prompts can significantly refine the capabilities of language models in solving intricate problems. This holds considerable promise for enhancing the pedagogical effectiveness of such technologies, paving the way for more sophisticated applications in educational domains where complex reasoning is required. Our ongoing research will continue to focus on these aspects, seeking to unlock further potentials and fine-tune the strategies outlined in this paper.

1 Introduction

In the field of mathematical education, enhancing students' problem-solving skills has always been a paramount goal. With the rapid advancement of artificial intelligence, particularly in natural language processing (NLP), large-scale predictive models such as those in the GPT series have begun to be integrated into educational technologies to assist in mathematical learning and improve problem-solving capabilities. However, despite their prowess in understanding and generating natural language, these models often struggle with mathematical word problems, which require logical reasoning and mathematical skills. This challenge arises because, while capable of capturing the surface structures of language, these models frequently overlook the deep logical and computational processes necessary for solving mathematical problems.

This paper explores the use of large predictive models in reasoning through mathematical word problems. We begin by defining what "reasoning" entails in the context of mathematical problem-solving and how it can be simulated in language models, particularly in aiding these models to understand and perform mathematical operations to derive correct answers. We introduce existing methodologies that leverage the intrinsic capabilities of these models, utilizing structured data input and enhanced training strategies to bolster their application in the mathematical domain.

Through an in-depth analysis of the performance and limitations of these predictive models in solving math problems, our aim is to propose a novel framework that integrates mathematical reasoning, problem decomposition, and solution verification. This framework is designed to improve the accuracy and reliability of models, making them more effective in educational settings. Moreover, we discuss how this approach can be adapted to different types of mathematical problems, ranging from simple arithmetic operations to more complex algebraic and geometric queries.

By appropriately training and tuning these models, we anticipate significant enhancements in their ability to solve mathematical word problems, thereby offering students a richer and more interactive learning experience. This paper also examines the implications of such technologies in personalized learning environments, where models can be tailored to address individual learning styles and needs.

Furthermore, we explore the ethical and practical considerations of deploying AI in educational contexts, highlighting the importance of transparency, fairness, and accountability in AI-driven educational tools. As these predictive models become more integrated into educational practices, it is crucial to address these aspects to ensure they contribute positively and equitably to educational outcomes.

Our research not only sheds light on the capabilities and challenges of using large predictive models in educational settings but also provides a comprehensive framework for enhancing their application in solving mathematical word problems. This contributes to the broader discourse on the intersection of AI and education, proposing pathways for future research and development in this exciting and rapidly evolving field.

In recent years, the application of large language models in solving mathematical word problems has shown promising advancements, particularly through the utilization of step-based reasoning prompts. These prompts decompose complex problems into manageable steps, enhancing the model's ability to perform logical reasoning. However, while this approach has effectively improved reasoning capabilities in mathematical contexts, it is not without its drawbacks. One significant challenge is the substantial increase in the length of both the prompts and the outputs required, whether in datasets used for supervised fine-tuning or in the assessment processes employed during testing. This escalation not only increases the computational and memory requirements but also inflates the overall cost of training and deploying these models

To address these issues, we propose a novel approach that emphasizes a purely numerical reasoning structure. By strategically integrating different formats of datasets for hybrid training, we aim to refine the efficiency of the model without sacrificing its performance. Specifically, we utilized the GSM8k and MATH datasets, which are pivotal in benchmarking models against current state-of-the-art (SOTA) algorithms. Our methodology involves contrasting these SOTA algorithms with our own, which leverages a more concise and numerically focused framework.

Our findings demonstrate that, by reducing the dependency on lengthy textual explanations and instead concentrating on numerical data, we can significantly decrease the number of tokens required for model operation. This reduction is achieved without substantial losses in performance, showcasing a balance between efficiency and effectiveness. Furthermore, this approach not only minimizes the computational load but also enhances the practical applicability of these models in real-world educational settings, where resources may be limited.

This paper delves into the technical specifics of our methodology, including the design and implementation of the numerical reasoning structure and the hybrid training process. We also discuss the implications of our findings for future research, particularly in the optimization of language models for educational purposes. By reducing the verbosity of model interactions, we pave the way for more accessible and scalable AI tools in education, potentially transforming how students engage with complex mathematical problems.

2 Related Work

In this section, we will introduce some of the methods we have employed.

2.1 Supervised Fine-Tuning

In the landscape of natural language processing, large pre-trained models such as GPT-3 and BERT have demonstrated remarkable capabilities in capturing rich linguistic features from extensive unlabelled datasets through self-supervised learning mechanisms. These models, developed using vast amounts of text data, learn to predict text or classify tokens in a way that encodes a deep understanding of language nuances, grammar, and context. However, despite their general efficacy across a broad spectrum of linguistic tasks, these models often require targeted refinement to excel in specific applications, a process known as Supervised Fine-Tuning (SFT).

Supervised Fine-Tuning (SFT) involves adapting a pre-trained model to a specific task by continuing the training process on a smaller, task-specific dataset that includes labeled examples. This process allows the model to hone in on the nuances and unique requirements of the task at hand. For instance, while a model pre-trained on a diverse corpus might possess a generalized understanding of language, it might not perform optimally on specialized tasks such as legal document analysis or biomedical text interpretation without fine-tuning.

The principle of SFT rests on the assumption that a small amount of targeted training data, when used to fine-tune a model that has already learned broad language representations, can significantly boost performance on specific tasks. The effectiveness of this approach has been demonstrated across various domains, including text classification, sentiment analysis, named entity recognition, and machine translation.

One of the critical challenges in SFT is avoiding overfitting, especially when the fine-tuning dataset is relatively small. Researchers have employed several strategies to mitigate this risk, including the use of regularization techniques, early stopping, and selectively fine-tuning only the top layers of the model while keeping lower layers frozen. These approaches help maintain the generalizability of the model while adapting it to specific tasks.

Additionally, the choice of hyperparameters such as learning rate and the number of training epochs plays a crucial role in the success of SFT. Too aggressive a learning rate can lead the model to converge too quickly to a suboptimal solution, while too many epochs can lead to overfitting on the training data.

Despite its successes, SFT is not without limitations. The dependence on labeled data for fine-tuning is a significant bottleneck, particularly in domains where such data is scarce or expensive to obtain. This has spurred interest in semi-supervised and unsupervised fine-tuning techniques, which aim to leverage unlabeled data alongside labeled data to improve model performance.

Looking forward, the integration of SFT with other advanced machine learning techniques such as meta-learning and reinforcement learning presents a promising avenue for research. These approaches could potentially enhance the model's ability to generalize from limited examples and adapt more efficiently to new tasks without extensive retraining.

Moreover, as the field advances, the exploration of more robust fine-tuning methodologies that require fewer data and computational resources will likely be a key focus. Addressing these challenges could broaden the applicability of SFT, making cutting-edge language models more accessible and useful across a wider array of tasks and industries.

2.2 Reformatted Alignment

In the burgeoning field of language model alignment, the introduction of the REALIGN method represents a significant stride towards refining the alignment of large language models (LLMs) with human values through enhanced instruction data quality.REALIGN (Fan et al., 2024) ingeniously addresses the pivotal challenges associated with the current instruction data enhancement techniques, which often involve labor-intensive processes or suffer from the propagation of factual inaccuracies due to LLM-generated content.

The REALIGN method is distinguished by its streamlined three-step process that minimizes the need for extensive manual annotation while significantly curbing the typical factual errors encountered in LLM outputs. The initial step in the REALIGN process is the Criteria Definition, where specific formatting criteria are established based on human-defined preferences across various scenarios. This step ensures that the instruction data not only aligns with human values but also meets structured and clear communication standards.

Following the Criteria Definition, the Retrieval Augmentation step extends the information base of the responses, particularly for tasks that demand high factual accuracy such as open-domain question answering and fact verification. By integrating relevant external information, this step enriches the depth and accuracy of the responses, enhancing their utility and reliability.

The final step, Reformatting, adjusts the original responses to conform to the predefined criteria and the augmented evidence base. This restructuring is crucial as it guarantees that the outputs are not only factual but also adhere to a consistent and clear format, thus improving readability and the overall quality of interaction with the LLM.

This approach contrasts sharply with prior methods that predominantly rely on either fully manual data creation or automated data extraction from existing datasets. While manual creation allows for high-quality, complex queries and responses, it is not scalable and is resource-intensive. Automated methods, on the other hand, often replicate errors and biases present in the source data. REALIGN's methodical enhancement of existing datasets through a defined, scalable, and largely automated process offers a novel solution that bridges the gap between these two extremes, paving the way for more sophisticated and aligned LLM interactions.

2.3 Dot by Dot

The research conducted by Pfau, Merrill, and Bowman(Pfau et al., 2024) delves into the computational dynamics of transformers within large language models (LLMs) by exploring the impact of intermediate, non-informational tokens—dubbed 'filler tokens'—on model performance. Their study is situated within the broader context of chain-of-thought processing, which has been increasingly recognized for enhancing model responses across various benchmarks.

This work contributes to the ongoing discussion on the expressivity and computational capabilities of transformers, particularly when they are not constrained to generate semantically meaningful intermediate outputs. The key question addressed is whether the performance improvements attributed to chain-of-thought reasoning are genuinely due to a more human-like task decomposition or merely the result of the additional computation facilitated by these intermediate tokens.

In their experiments, Pfau and colleagues demonstrate that transformers can utilize sequences of meaningless filler tokens (e.g., sequences of dots) to achieve computational tasks that are otherwise unattainable without these additional tokens. This finding challenges the conventional understanding of token utility in transformers and suggests that the computational benefits of additional tokens may be independent of their semantic content.

Their analysis extends into a theoretical characterization of the types of problems where filler tokens are advantageous. This is particularly relevant for understanding the limitations and potential of transformers in processing complex computational tasks that go beyond simple next-token prediction.

This line of inquiry not only sheds light on the fundamental aspects of transformer architecture but also prompts a reevaluation of how transformers are trained and utilized in natural language processing tasks. It underscores the potential for transformers to engage in what the authors term "hidden computation," where the computational work done by the model does not transparently align with the generated tokens.

This research sets the stage for further explorations into the design and training of LLMs, emphasizing the need for more nuanced approaches that consider the underlying computational processes rather than just the observable outputs. This could lead to more efficient models that better leverage the inherent capabilities of transformers, potentially transforming practices in AI model training and deployment.

3 Method

In this section, we will introduce the main improvements we have made to the methods and propose a new evaluation criterion that considers both accuracy and the average number of tokens used.

3.1 Focusing On Numbers (FoN)

In this study, we conducted an in-depth analysis of the inherent issues within the GSM8k dataset, which is predominantly comprised of problems that resemble mathematical word problems. Traditional state-of-the-art (SOTA) algorithms typically engage in step reasoning that involves a substantial use of tokens to delve into aspects such as background analysis and character relationships. However, we observed that these elements, while providing narrative context, do not significantly aid in arriving at the final mathematical solutions.

To address this inefficiency, we developed an innovative algorithm named Focusing On Numbers (FoN). The core philosophy behind FoN is to streamline the reasoning process by concentrating on the essential quantitative elements of the problems—specifically, the numbers and formulas that are crucial for problem-solving. By strategically omitting extraneous textual information, our algorithm effectively reduces the number of tokens utilized in the reasoning steps. This approach not only enhances the efficiency of the computational process but also potentially improves the overall performance of the model by focusing on the most critical data points necessary for mathematical reasoning.

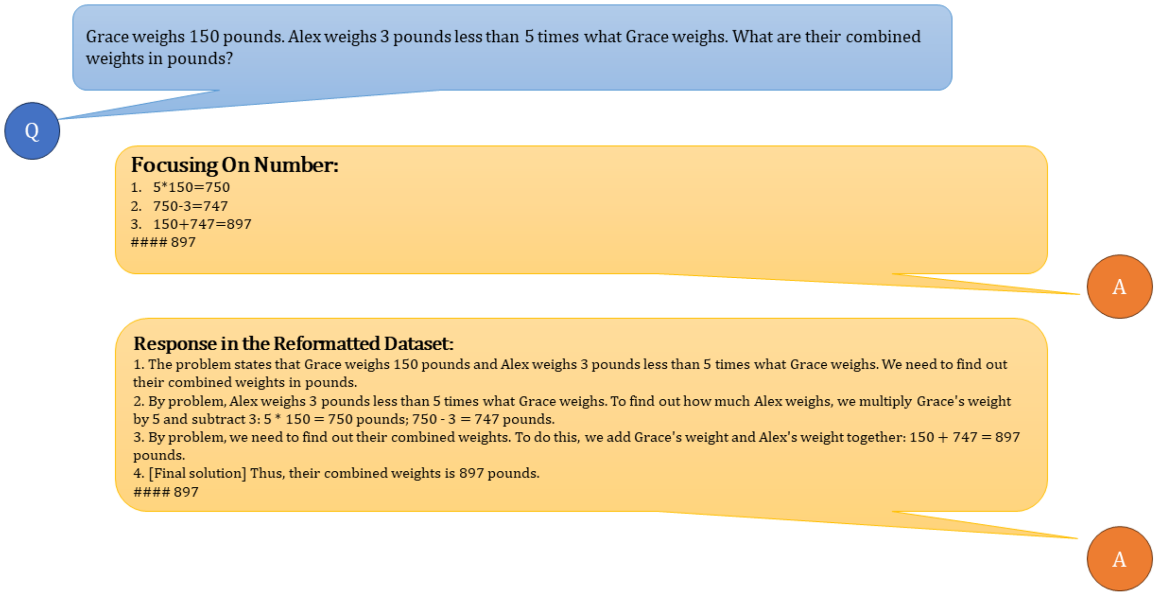

Figure 1: FoN Demonstration.

3.2 Data Augmentation through Greater Models

We used gpt-3.5-turbo to regenerate answers with fewer steps. Question+' Answer this question in $\frac{n}{2}$ steps' where n is the number of reasoning steps of response to this question in the Reformatted Dataset.

3.3 Hybird Training

In our methodology, we employed a hybrid training approach during the supervised fine-tuning of our large predictive models. This involved mixing three different dataset formats at varying ratios: "Only Answer," "Full Text," and "Numbers Only." Each dataset serves a specific purpose:

Only Answer: Focuses on direct numerical answers to enhance the model's precision.

Full Text: Provides detailed problem descriptions and solutions to improve comprehensive textual understanding and reasoning.

Numbers Only: Concentrates on essential calculations to boost numerical reasoning and computational efficiency.
0

0 By blending these datasets according to specific probabilities, we aimed to balance the training emphasis between textual comprehension and numerical accuracy, thus enhancing the model's adaptability and generalization across various mathematical problems. This strategy allows for a focused evaluation of how each dataset type influences model performance in terms of accuracy and efficiency.

3.4 New Score Function with Respect to Efficiency

Previously, evaluations of the reasoning capabilities of large predictive models were confined solely to the accuracy of the results. Recognizing the importance of efficiency in reasoning, we have also incorporated the average number of tokens used by the model to complete each reasoning task into our evaluation criteria. Moreover, we ensured a prerequisite: it is only meaningful to consider the efficiency of reasoning if the reasoning results are correct.

Therefore, we devised the evaluation function as (1)

$$score = \sum_{i=0}^{n} \mathbf{1}\{output_i = y_i\} \cdot (\frac{L_{baseline} - L}{L_{baseline}} + 1) \tag{1}$$

where n is the size of dataset and $L_{baseline}$ presents the average token number used by baseline (Reformatted Alignment), L presents the number of average tokens used by the method to score.

Hyperparameters name	Value
Batch size per device (training)	8
Gradient accumulation steps	8
learning rate	2×10^{-5}
learning epochs	3

Table 1: Hyperparameters settings

4 Experiments

4.1 Experiment Setup

We chose the Mistral-7b-v0.1 (Jiang et al., 2023) as our base model, and demonstrated SFT by GSM8k dataset and our specified datasets. We tested the accuracy of those tuned models on various benchmarks, GSM8k, GSM8k Robust and MATH. In the context of efficient reasoning, we also rolled out the token lengths of their answers. Generally, the more accuracy and fewer token occupied, the better efficiency in reasoning. For SFT we adopted the supervised fine-tuning mode in Safe-RLHF structure (Ji et al., 2023) (Dai et al., 2024). Also, we used 4*NVIDIA® A800 40GB Active GPUs for SFT and 1 for evaluation, and we used Deepspeed structure for single node acceleration. The training time on average is approximately 1 hour. More experimenting details are in 1.

As for baselines, we demonstrated several customed Datasets: GSM8k, GSM8k w/ final answers, GSM8k + GSM8k w/ final answers, GSM8k w/ '...' + final answers. Those with direct final answers replace their responses with Simply the final numeral answers, which definitely has the least redundancy, but of course, it's way too hard to learn from those data. Those models tuned with final-answers-only data, in general, would take fewer reasoning steps to make an answer but so low were their accuracy, as can be seen in 2.

4.2 Metrics

We rolled out the accuracy of the models on GSM8k, MATH, and GSM8k Robust test Dataset. Also, we calculated the output token lengths of the answers of the models. We use the score metric mentioned in 3.4, and note that the GSM8k Score is derived from the average performance over GSM8k and GSM8k-Robust. Actually, the performances on these two datasets are too similar.

4.3 Results

We illustrated all the models' responses to the same question to clearly display the answering patterns of the models. The two direct baseline rollouts are in 2 (GSM8k and Reformatted) and you can feel how redundant their answers are. The focusing-on-answer model output is in 3, showcasing a totally different pattern that directly gives an answer, but the correctness is hardly ensured.

4.3.1 Focusing on Numbers

Dataset	GSM8k	MATH	GSM8k Robust	GSM8k Len	MATH Len	GSM8k Score	MATH Score
GSM8k (baseline)	47.61	3.84	46.02	53.95	81.95	46.81	3.84
GSM8k Final Ans	13.04	5.22	9.86	3.00	3.66	22.26	10.21
GSM8k '...'+Final Ans	10.6	5.46	10.08	4.00	5.16	19.91	**10.61**
GSM8k FoN(Ours)	**50.87**	3.56	**48.52**	9.60	8.82	**90.55**	6.74

Table 2: FoN versus other baselines on original GSM8k Dataset. Note that the GSM8k Token Len refers to the average token lengths over GSM8k and GSM8k-Robust; and the MATH Token Len refers to those of MATH Dataset. The same in 3.

From Table 2, we can observe that the FoN(Ours) technique attains a dominant performance against all the baseline, leveraging a relatively fewer reasoning lengths with almost no loss in accuracy. We can see that FoN even has better accuracy than GSM8k Vanilla. For instance in 4, it just derives the formulas like most of the grade school students do their maths, and gives the answer.

So in this case, FoN greatly reduces redundancy with seldom loss in accuracy. But FoN does not work so well for the Reformatted Dataset, especially on MATH evaluation, as can be seen in 3. This is remarkably questioning because we wanted the model to focus on numbers, but the model did so bad in MATH.

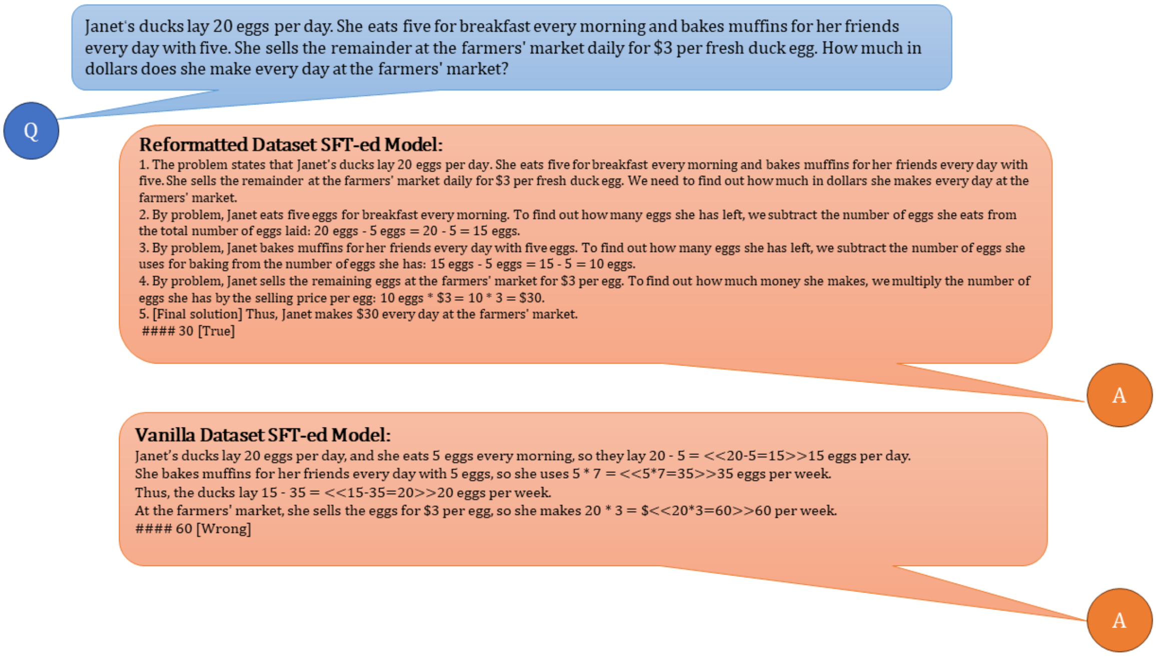

Figure 2: Responses of the baselines GSM8k Vanilla and GSM8k Reformatted

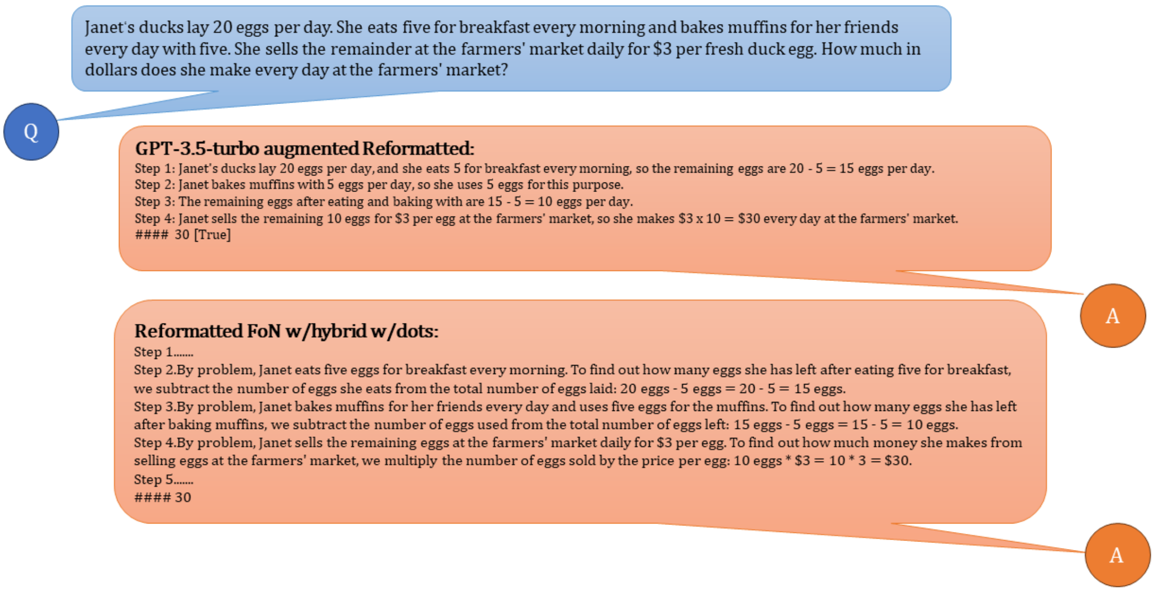

Figure 3: FoN w/ hybrid w/dots and Data Augmented Responses

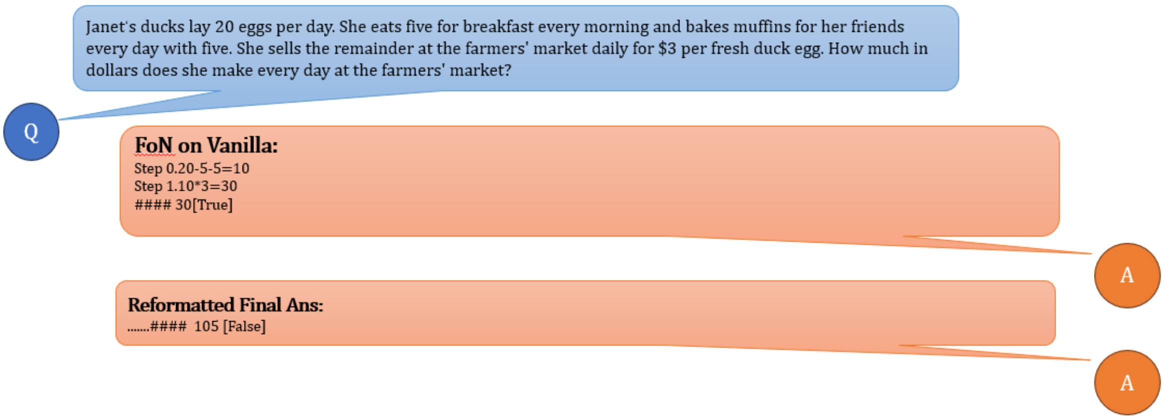

Figure 4: Inference responses of FoN on vanilla, Focusing on Answer on Reformatted models.

Dataset	GSM8k	MATH	GSM8k Robust	GSM8k Token Len	MATH Token Len	GSM8k Score	MATH Score
GSM8k Reformatted	**61.26**	4.80	**58.91**	219.22	328.66	60.08	4.80
GSM8k Reformatted + GSM8k Reformatted w/ Final Ans	11.07	4.72	9.90	3.00	3.84	20.83	9.38
GSM8k Reformatted FoN (Ours)	47.31	3.30	46.10	42.32	130.74	**79.39**	7.91
GSM8k Reformatted FoN w/ hybrid, w/o dots	43.67	3.54	42.91	126.58	55.81	61.58	6.48
GSM8k Reformatted FoN w/ hybrid, w/ dots (Ours)	53.83	5.32	52.77	148.83	87.25	70.41	9.23
GSM8k Reformatted Augmented (Ours)	42.15	**6.42**	38.36	72.46	144.99	67.20	**10.00**

Table 3: FoN w/ hybrid&dots, Augmentation versus other baselines on reformatted GSM8k Reformatted Dataset

4.3.2 Hybrid

In result, we developed some tricks to improve its performance on GSM8k. One is to hybrid. For each reasoning step in the Reformatted Dataset, we have three options: to retain the numeral formula, fully reserved and replaced by dots('......'). These three options has its distinctive advantages and shortages: retaining numerals helps the model focus on numeral logic, but suffers from huge information loss; replacing by dots has more severe shortage than reserving numerals, but it's the best way for removing redundant information; fully reserving has no information loss, but is the most redundant. To sum up, we shall hybrid the three methods, so we developed a probabilistic choosing strategy. We tried different fixed probabilities:

- If a step has a formula: p_1 for fully reserved, $1 - p_1$ for reserving numeral formulas. Zero probability for dots because those steps with formulas are regarded as important, so
- If a step has no formulas: p_2 for fully reserved, $1 - p_2$ for replacing with dots('......').

After multiple trials in 4, we observed that $p_1 = 0.8, p_2 = 0$ got the top results, and that's what 'w/hybrid' in 3 refers to. We also did ablation on either replacing dots with non-formula steps. 'W/o dots' means we simply discard the non-formula steps, and it suffers a remarkable loss in scores. So, as seen in 3 and 3, SFT w/ hybrid w/ dots dataset achieved a competitive performance, reducing output token lengths and gives a relatively more concise response.. It finally improves MATH accuracy as we expected.

4.3.3 Data Augmentation through Greater Models

For augmentation, we derived more simplified answers from gpt-3.5-turbo by adding a suffix to the questions to ask the greater models to output answers with fewer reasoning steps. In the experiment, we asked gpt-3.5-turbo to respond within half of the reasoning steps in the reformatted dataset and collected its responses as a new dataset for augmentation. As seen in 3, this method achieved considerable improvements in MATH evaluations. And for

p_1 (full/formula)	p_2 (full/dots)	GSM8k	MATH	GSM8k Robust	GSM8k Token Len	MATH Token Len	GSM8k Score	MATH Score
0.0	0.8	47.31	3.30	46.30	42.32	130.24	84.57	5.29
0.2	0.0	45.41	4.86	45.87	69.02	49.25	69.91	8.99
0.5	0.0	39.50	2.76	36.85	90.12	53.21	60.66	5.07

Table 4: Multiple trials on probability choosing in Hybrid

instance, in 4, data augmentation further simplified the response. This is because the answers from greater models might have a better representation on numbers and math logic than our arbitrarily picking the numeral formulas up. Its loss in GSM8k performance might be attributed to the radical change in its GSM8k training dataset, while others use neatly modified datasets that have better similarities with the GSM8k testing dataset.

5 Conclusion

In our research, we devised a groundbreaking strategy aimed at enabling large predictive models to efficiently engage in reasoning to solve mathematical problems while conserving the number of tokens used, an approach we have termed ”Focusing on Numbers.” By analyzing the actual structure of mathematical questions, this method prioritizes numerical data and critical computational steps, significantly reducing token expenditure without sacrificing the depth of reasoning required for complex mathematical tasks.

Additionally, we introduced a novel evaluation system specifically tailored to assess these large predictive models in terms of both accuracy and efficiency in solving mathematical problems. This dual-focused evaluation framework is essential in quantifying the effectiveness of the models not just in reaching the correct answers but also in how economically they utilize computational resources.

Using this new assessment method, we evaluated existing state-of-the-art (SOTA) algorithms alongside our ”Focusing on Numbers” strategy. We further tested the efficacy of using enhanced language models for data regeneration and the impact of training with a hybrid of different structured datasets. Our results demonstrate that our approach not only competes with but in some cases surpasses, the performance of current SOTA algorithms in terms of efficiency and accuracy.

The evaluation revealed that focusing on numerical data allows for a more streamlined model operation, which is particularly advantageous in educational settings where computational efficiency is paramount. Moreover, the regeneration of datasets with stronger language models and the strategic hybrid training approach have shown that our method can adapt and excel even under varied and complex problem sets.

In conclusion, our research contributes significantly to the field of AI in education by proposing an efficient, accurate, and token-economical approach to solving mathematical problems. The ”Focusing on Numbers” strategy, complemented by our innovative evaluation framework, sets a new benchmark for future research in enhancing the capabilities of AI models in educational technology. As we move forward, these findings could pave the way for more sustainable, effective, and accessible AI-driven educational tools, potentially transforming how learners engage with and overcome mathematical challenges.

6 Future Work

In our experiments, we only tried some fixed hybrid probabilities; this is kind of too arbitrary and insufficient. We propose that the probability distribution of selecting simplification methods can be modeled like Gaussian, etc., guided by the data. This is quite an interesting and feasible orientation for further research.

Acknowledgements

Thanks to the guidance of Prof. Pengfei Liu, Shijie Xia and computing resources provided by GAIR lab.

References

[1] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

[2] Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. Reformatted alignment. *arXiv preprint arXiv:2402.12219*, 2024. URL https://arxiv.org/abs/2402.12219.

[3] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=g0QovXbFw3.

[4] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[5] Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.