

# **Fine-Tuning Open-Source Large Language Models for Generating Math Explanations**

Neena Xiang, Paul Godinez, Eli Hoffberg

## **Acknowledgements**

We would like to express our gratitude to Professor Neil Heffernan for advising and guiding our project, as well as to PhD Student Eamon Worden for his guidance and oversight over the course of our project. We would also like to thank ASSISTments Grant Manager Angela Kao for coordinating meetings and providing insight into the ASSISTments program. We would like to express our gratitude for the ASSISTments team and share our respect for the ASSISTments program as a whole for working with us and providing us with data. Lastly, we would like to express thanks to the WPI Computer Science Undergraduate Program for providing us with the opportunity and funding to complete our project.

**"This report represents the work of one or more WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on the web without editorial or peer review."**

## **Abstract**

Percy Liang’s article, “We have No Moat,” reveals that open-source large language models (LLMs) with 7 billion parameters are able to rival those of large tech companies with 500 billion parameters. Open-source LLMs have also become more accessible and easier to fine-tune with the rise of open-source resources like Hugging Face. Through the use of prompt engineering and fine-tuning, the goal of this project was to find and evaluate LLMs to potentially match the performance of OpenAI’s GPT-3.5. We aim to help ASSISTments, a non-profit organization that focuses on middle-school math education, in developing open-source LLMs to transition from tedious and somewhat inaccurate hand-written explanations to streamlined automatically generated ones. Open source LLMs offer a more cost-effective option compared to GPT-3.5 and a more time-efficient option compared to generating explanations by hand. ASSISTments has already started working on integrating LLMs into their website, and our focus was on improving the explanation generating LLMs.

Leveraging a framework of prompt engineering and fine-tuning LLMs, we tested and evaluated the effectiveness of many models in writing accurate math explanations. During prompt engineering, we double-blinded the responses for each prompt and evaluated each response. This double-blind process allowed us to determine the score in an unbiased manner. Through an iterative process, we were able to see up to 80% improvement with our best prompts compared to just giving a labeled question-answer pair to prompt the LLM. Performing fine-tuning, we determined that we were unable to significantly improve a WizardMath’s mathematical reasoning, but fine-tuning was highly effective in producing consistently formatted answers which gave the explanations more readability compared to the base WizardMath. This framework was ultimately used to compare the performance of 3 LLMs in generating explanations to ASSISTments questions. We found that the fine-tuned model improved the base model by about 5%, while GPT-3.5 outperformed the base model by roughly 45%. Our results show promise in utilizing LLMs for generating accurate and readable explanations. Furthermore, our fine-tuning and prompt engineering framework can be utilized in other fields in which LLMs can be integrated in order to optimize the performance of the LLMs.

## **1 Intro**

ASSISTments is a non-profit organization created to help teachers teach math to students. Its website is designed to help students in a way that keeps them engaged by giving them more challenging problems when they succeed, and by providing just enough hints to let them work out the problem when they’re struggling. It also tabulates the data for teachers so that they can better curate their lesson plans for when teaching. The point of our study is to determine the possibility of using large language models to improve the ASSISTments platform, specifically when explaining math problems to students.

Large language models, or LLMs, have been becoming more widespread in a multitude of industries. They can be used as chatbots to provide information to employees and customers,

and as assistants for tasks like computer programming, writing, and brainstorming (Metz, 2023). For example, Khan Academy recently deployed Khanmigo, an LLM chatbot for the purposes of education (Khan Academy, 2024). Some of Khanmigo's use cases as listed on its official website include grading papers, creating lesson plans, and guiding users through topics without giving specific answers. However, the LLM that was at the forefront of bringing LLMs to the public eye is OpenAI's GPT. OpenAI's GPT is not open source (OpenAI, Achiam, Adler, et al., 2023), but other popular LLMs that are open-source include Meta's Llama (Meta, 2023) and Google's Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Ming-Wei, 2023).

The major difficulty in utilizing LLMs to solve math problems is that they generate responses based on their training data. This means that when an LLM is given a math problem, instead of making calculations, it will predict an answer based on solutions to math problems it has seen in the past. There are few LLMs that specialize in math questions because in general, LLMs are not designed to calculate and solve math problems. The WizardLM team was one of the few groups to address this issue with their WizardMath LLM, using their own framework to make LLMs understand and parse math questions (Luo, Sun, Xu, et al, 2023). Our project aims to improve LLMs for solving math problems through both fine-tuning models with significantly lower parameter counts than GPT-4's 1.7 trillion parameters (OpenAI, et al., 2023) and engineering prompts by adding onto and combining existing frameworks for those models to get the most coherent and correct math explanations possible.

## **2 Background**

### **2.1 Fine-Tuning LLMs**

One of the greatest advancements in LLMs is the availability of open-source LLMs. Access to models has been easier than ever with websites, such as Hugging Face, storing and sharing original or fine-tuned open-source LLMs (Jain, 2022). Additionally, the barrier of entry to tinker with these LLMs has never been lower (Patel & Ahmad, 2023). Fine-tuning broader models to fit individual needs has eliminated the need to train models from scratch. As a result, models need much less time and money to train and can even be trained on a personal device, as fine-tuning significantly lowers the required computational power. Due to the availability of high-quality open-source LLMs, fewer resources are required to produce an LLM of comparable quality to private LLMs. Even Google believes that open-source AI will outperform those of private corporations, such as those from OpenAI and Google themselves.

A major development in fine-tuning came with the implementation of low rank adaptation (LoRA) which allows fine-tuning at a fraction of the cost and time (Hu, Shen, Wallis, et al., 2021). LoRA allows for fine-tuned adjustments to stack, eliminating the need to run a model from scratch or retrain. It also enables personal devices with modest computing resources to customize models quickly, which makes fine-tuning LLMs significantly more accessible. For context, a previous requirement for fine tuning was access to a powerful server or machine. Patel and Ahmad (2023) in their leaked memo about open-source AI also note that larger models

aren't strictly better than smaller ones and come with two major cons: less adaptability and longer training time. Therefore, this project leverages the power of LoRA to focus on fine-tuning smaller models.

The first research group to successfully fine-tune an LLM to outperform the base model was Alpaca (Taori, Gulrajani, Zhang, et al., 2023). Alpaca 7B (7-billion-parameter) is an open-source LLM that is derivative of Meta's LAmMA 7B. A team at Stanford was able to create their model easily and cheaply, spending less than \$600. They found that the 2 main challenges they faced were finding a strong base LLM and collecting high-quality data. Taori et al. were able to generate 53 thousand data points by instructing a different model to create additional instructions based on 175 human-written instructions. Alpaca 7B worked well but still suffers from the same deficiencies that most LLMs suffer from inaccurate responses, toxicities, and stereotypes. They also found that the final model shows the data deficiencies from the original model and the fine-tuning data. This project kept those points in mind as we strove to create a model that is suitable for math tutoring. Additionally, we utilized and improved upon their training recipe as a template for our own process.

## 2.2 Models of Interest

This project involved the use of open-source LLMs for data collection and potential use for ASSISTments multistep explanation generation, including Mistral, Llemma, and WizardMath.

Mistral is a relatively new open-source LLM created by the French company, Mistral AI (Jiang, Sablayrolles, Mensch, et al., 2023). The company gives users the ability to fine-tune their Mistral models in order to solve specific issues, and this widespread free use helps the developers find more bugs and improve the base Mistral model (Metz, 2023). Mistral was chosen as a model of interest for this project due to its ability to generate answers to questions, and its developer and community support. Another new open-source LLM used as a focus for this project was Llemma, which is focused solely on using generative AI to solve math problems (Azerbayev, Schoelkopf, Paster, et al., 2023). While models with more input parameters seem to work better in terms of correctly and concisely generating responses (Azerbayev et al., 2023), we utilized Llemma and Mistral at 7 billion parameters due to the speed and ease of training each model. Unlike smaller models such as Llemma 7B and Mistral 7B, models with more parameters require more space to store, can take longer to fine-tune and run prompts on, and can be more expensive to run and store (Fathi, 2023).

Similar to Mistral and Llemma, WizardMath is an LLM that was released recently. Built on top of a preexisting LLM named WizardLM, the Wizard series of LLMs is developed by the WizardLM team who is partnered with Microsoft (WizardLM, 2023). Their goal is to "build AGI for humanity and the world," (WizardLM, 2023) with an emphasis on keeping LLMs open-source, and utilizing the pre-training step for LLMs. With our team's focus on open-source models and Math explanations, WizardMath appeared to be the perfect fit for our team's purposes, and once we had discovered it, we quickly decided to make it our base model to

fine-tune on. The WizardMath series was developed by the WizardLM team's own three step Reinforcement Learning from Evol-Instruct Feedback (*RLEIF*) method, which includes their own steps for fine-tuning and reward programs for making LLMs understand math problems and equations (Luo, et al, 2023). We believed that we could iterate on this framework to make LLMs more likely to explain the steps to the math problems it may be given. When our team discovered WizardMath, we decided to utilize both the 7B and 13B (13-billion-parameter) models, one for prompt engineering, and one for fine-tuning. We hoped that we would be able to save time running models by prompt-engineering on the 7B, and improve our responses by using these prompts on the 13B model.

### **2.3 Prompt Engineering**

Prompt engineering is a field that has become more popular as LLMs have become more complex. With increased complexity, getting an LLM to return a desired result becomes more challenging. Methods of prompt engineering include changing certain variables in input parameters, such as booleans that cause results to be less repetitive and integers that increase the number of characters in a response or increase the randomness in a response. Prompt engineering can also include finding the best word choice for an LLM to produce the best possible response. Due to prompt engineering being a new field of study, there are not many standards for executing it, with the one of the most recent papers describing such a standard being only released in 2022 (Wei,, Wang, Schurrmans, et al., 2022).

In addition to fine-tuning, our team realized early on that we would need to work on prompt engineering as a large portion of our methods. We would need to fine-tune models, and at the same time, find the prompts that would work best for those models and compare the responses. We build on *EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria* (Kim, Lee, Shin, et al, 2023) which heavily influenced our prompt engineering cycle. We were inspired by this paper to use different prompts for one instruction so that we may compare the two, but decided to add our own twist and single blinded the participants who were grading the prompts so that we may have a fairer grading process.

## **3 Method**

We gathered data from a variety of sources which include, the ASSISTments database, GPT-4 generated, and human-generated explanations. The data was then used to fine-tune models to meet the goal of writing explanations. Those models that were fine-tuned were then compared to their original counterparts and rated based on how accurate and clear they were. The following section will explain our decision-making and direction in building a framework that would result in a suitable LLM that could be implemented in ASSISTments.

### **3.1 Data Acquisition**

The original plan to gather data was to use the ASSISTments database to gather a set of teacher-written problem explanations. Looking at that data, we found that there were many

formatting issues when the data was taken from ASSISTments. We cleaned the data by removing all the HTML tags and entities that were in the responses to ensure that any LLMs fine-tuned with this data wouldn't print any tags or entities. Any personally identifiable information was also removed to ensure the models would not learn any personal information about ASSISTments teachers or students.

Beyond the initial formatting problems in the data, we discovered that there was a bigger issue with the data quality. Some of the explanations written by teachers contained inaccuracies that could negatively influence the accuracy of the final models. The teacher-written explanation formats given were also inconsistent. Consequently, we decided to pivot to a synthetic dataset that was generated by OpenAI's GPT-4. There were 2 main advantages to this method of data collection i) more consistent response explanations and ii) more accurate data. Step-by-step explanations were generated for the 918 ASSISTments questions provided for our training set.

To test the explanation models, we created 2 datasets for different purposes. The first test dataset we created was for prompt engineering. We wrote 20 seventh-grade level questions from scratch to ensure the LLMs would have to answer a question they have never seen before. This test set was smaller because we tested more models with more prompts that were reviewed by hand. The second dataset was created to test the performance of the fine-tuned models against GPT-3.5 specifically on ASSISTments questions. Therefore, this dataset consisted of 50 questions and answers that were curated from questions that were excluded from the training set. The models tested with this dataset would have to write explanations with a provided question and answer pair. The ASSISTments database contains the corresponding answers to each question, so providing the LLMs with the answer to the question should increase the accuracy of the explanations. This final test set would provide more data points when testing the statistical significance of the model improvement.

### **3.2 Fine Tuning and Prompt Engineering Cycle**

To ensure greater success when fine-tuning a model, we needed to find the best base models to work off. We wanted the models to have stronger math and logic capabilities to properly explain or give feedback for math problems, so we referenced Kumar's (2023) article of his top 13 math LLMs when finding a good base model. Additionally, we looked at the Hugging Face website's LLM leaderboard which ranked LLMs by various metrics. The metric that we were most interested in was TruthfulQA because it would translate into more accurate responses. We found that 3 open-source LLMs stood out at the time: Mistral, WizardMath, and Llemma. With the LLMs selected, they would be fine-tuned on the dataset that was specific to the task of either writing explanations or providing feedback.

After selecting the models to work with, we started prompt engineering the base models while fine-tuning them at the same time. Once the models were fine-tuned, they would join their base models in the prompt engineering process. The goal of prompt engineering was to find what types of prompts would have each model produce the best response. As mentioned earlier, our framework for prompt engineering was inspired by the *EvalLLM* paper. In their paper, they

describe a framework where they take the same question and use two different prompts on the same LLM. They would then grade the responses the LLM gave based on a set of criteria they had made. Our framework began the same way as the framework outlined in the *EvalLM* paper, but we expanded it further to include blinding. Our team did this to make prompt engineering more of a science instead of a guess and check for each prompt.

We began by having two researchers prompt two different LLMs, using multiple prompts and gathering answers based on a set of questions for each prompt. Similar to *EvalLM* these two researchers would then agree on a set of criteria so that they may grade each response for the prompt on a scale from 0-4. However, this is where our team's framework diverged from *EvalLM*, as each researcher would swap datasets with the other researcher, but be blinded as to which answers corresponded to which prompt. After grading it all, the two researchers would show each other which prompt corresponded to which grade they gave. They would then discuss the prompts with the highest scores and further iterate on them so that at the end of the prompt engineering cycle, they would be left with the best prompts for their LLMs.

As a final test of the effects from our fine-tuning on WizardMath, we ran GPT-3.5 with some of the common prompts we were using with the other LLMs. In this test, the larger, more robust dataset was used to test the difference in model accuracy. GPT-3.5 would still be graded with the same method with scores ranging from 0 to 4. Additionally, we had 2 members of our team grade it to evaluate if the grader made a difference in score.

## **4 Experiment**

### **4.1 Fine Tuning Process**

We fine-tuned multiple models to test how much of an improvement we could make from the base models. Initially, we turned the models on Google Colab and CPU-only, but faced issues with limited RAM and couldn't load any models that were too large. As such, we moved the fine-tuning to the WPI Turing Clusters which provided the resources needed to effectively fine-tune the models. Models were fine-tuned with Hugging Face's PEFT (Parameter-Efficient Fine-Tuning) library which allowed us to use LoRA to adapt the models to our specific task of giving clear problem explanations.

We adapted Alvarez's (2023) code that fine-tuned a small Vicuna model to suit our needs in terms of model and output format. Because of the length of the fine-tuning and testing process, we were not able to extensively explore the most optimal set hyperparameters, so we used the default ones that were provided in Alvarez's article. We utilized a learning rate of  $2e-4$  with a batch size of 4 and a dropout rate of 0.1. The max step size for tuning the initial models was 1000, but this was increased to 2000 when tuning the final WizardMath model that would be compared to GPT-3.5. Future work would include allocating additional time for model training to improve results, as the current input parameters, which were set to allow adequate time to test the model, limited the training time to 19 hours.

Initially, we were testing the models' math and logic ability, so they were trained only on a given question and explanation. These models were used to find the best prompts and

commonalities between the successful prompts. However, given that the ASSISTments questions had answers provided, the final model that would compare to GPT-3.5 would also be trained on a dataset that provided the question, explanation, and answer. The goal was to test the model’s logical reasoning to provide a clear and concise answer without having to guess the answer. The final model with trained with this prompt:

Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>: [Question] <Answer>: [Answer].

The same prompt was used for GPT-3.5 but had to be adapted for the base WizardMath model due to the formatting of the answer given this prompt.

## 4.2 Prompt and Model Evaluation

We worked with 3 main models and their fine-tuned counterparts: WizardMath, Mistral, and Llemma. To evaluate the models during the initial rounds of prompt engineering, each model was given multiple prompts and fed 20 manually-created test set questions to write an explanation for. These models then produced explanations that would be stored in a text format (CSV file). After originally testing with 256 tokens, we found that while responses were returned after 10 to 15 minutes in our implementation, the models were often unable to complete the explanation. As such, we increased the token count to 512 which allowed most explanations to finish. The rater who tested the model would then shuffle the columns and remove the respective prompts for another team member to grade for accuracy. The guideline for grading can be seen in Table 1. The average score of each prompt was then compared to see which types of prompts generally performed better than others.

Table 1: Guideline used to grade how well LLMs wrote an explanation.

Score	Meaning
0	Nonsensical or irrelevant information
1	Mentions math terms, unclear and wrong
2	Comprehensible but wrong and/or very incomplete
3	Almost right path/answer
4	Nearly perfect, right answer, right explanation

For the final test of comparing WizardMath, fine-tuned WizardMath, and GPT-3.5, all 3 models were tested with the 50 question test set by 2 team members. The same scoring system was used for this final evaluation. We ran paired t tests on the scores to determine if there was any statistical difference between the model performances. We also utilized an ANOVO test between all 3 models, and 2 paired t-tests (one between the base and fine-tuned model and another between the fine-tuned model and GPT-3.5). Additionally, we created histograms to see



the distribution of scores, as well as a confusion matrix between fine-tuned WizardMath and GPT-3.5 to characterize the difference between the models. We also ran a Cohens Kappa so see if there was any significant difference in the scoring of our two team members.

## 5 Results

From the initial tests with prompt engineering, we found that some prompts and some models performed better than others on average. While Llemma could be improved by using prompt engineering, utilizing a control with no prompting aside from the question and answer consistently resulted in incomprehensible responses which mostly received scores of 0. However, as we continued to cycle with prompts, we were able to decrease the total number of 0 responses, as shown in the Appendix, Figure 8.2 With our first prompt, Llemma had nearly 70% improvement over our control output with just the math question and answer, and our second prompt had about 45% improvement over our first prompt as shown in Table 2. WizardMath, on the other hand, had less than 50% accuracy without prompting, but was increased by 6% to just above 50% accuracy when using prompts. While a 70% increase in accuracy, followed by an additional 45% increase in accuracy seems like it would lead to a better overall accuracy, Llemma’s base model was still significantly worse overall than that of WizardMath, and we decided WizardMath would be more worth fine-tuning to see if response structure could improve our results.

Table 2: Score Improvements for Llemma Prompt-Engineering

Llemma	Control	Cycle 1	Cycle 2
<b>Evaluator 1:</b>			
Average Score:	0.154	0.462	0.846
Improvement (%):	N/A	66.7%	45.5%

While using prompt engineering to improve responses from the LLMs, we found that the most successful prompts had 3 common traits. They were generally more concise, didn’t use any negative language, and were well-tailored to the types of questions that were asked. Removing irrelevant information from the prompts helped the LLMs not to be confused when answering them. Avoiding negative language allowed for the LLM to not be confused when reading a statement in a prompt with a negative, and still perform the specifically stated undesired action. For example, telling an LLM to “stop repeating” itself when responding could cause it to see “repeat” and repeat itself in even more responses than before. Tailoring prompts to the types of questions asked would help the LLM focus on certain word associations and help it find the desired solution to a problem. A prompt we gave to Mistral that did not follow these guidelines was “Generate a correct answer to the following math problem”. This prompt did not ask Mistral to solve or explain the solution to a math problem, which is what we wanted, and resulted in responses that only gave answers without explaining them or attempted to use sources outside of itself with lines such as:

“Report an Error Ask A Question Download App”, and “[buy\_now\_button buynowlink=”https://expertwritingservices.net/shop/instruction-create-a-step-by-step-explanation-to-the-following-problem-person-a-gets-to-a-””.

This specific prompt resulted in scores 66% of that of the control. These discoveries then influenced how we prompted the final test of the models comparing our fine-tuned open-source model to GPT-3.5.

When fine-tuning the final WizardMath model, the 2000 steps that it was trained for translated to approximately 34.78 epochs. The initial loss was 0.8324 and the final loss after training was 0.0233. After the model was fine-tuned it was tested against the base WizardMath model and GPT-3.5. Between 2 graders the resulting score distribution can be seen in the histogram in Figure 1. We saw that most scores, which was around half the questions, were 4s for both graders which meant that the model was able to answer completely accurately about half the time. There were also only a few questions that received a score of 0 which means most explanations were coherent.

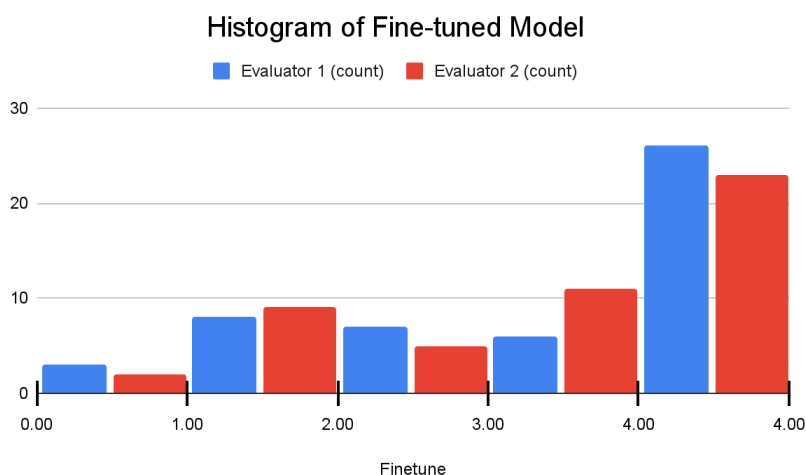


Figure 1: Histograms of both evaluators’ grading of the fine-tuned WizardMath model.

A Cohens Kappa score was calculated for all three models to test the agreement between the graders which can be seen in Table 8.1 in the Appendix. The resulting scores found that there was fair agreement between the two graders, as all the scores were between 0.2 and 0.4. Looking at the average scores and standard deviation of the models between both graders as seen in Table 3. We saw that the average for the fine-tuned model was the same at 2.88 which is greater than the averages given to the base model, showing a slight improvement. However, models averaged at least a whole point lower than GPT-3.5, which performed the best by far out of the models. It was also the most consistent with the smallest standard deviation across both graders. These numbers show about a 5% improvement from the base to fine-tuned model with high standard

deviation. GPT-3.5 is shown to outperform the base model by roughly 45% and the fine-tuned model by roughly 37%, as it is reflected in the averages and standard deviations.

After running statistical tests which can be seen in Table 4, we can see in the ANOVA test that there was one model that was statistically different between the 3 models since the p-value was less than 0.05. Running paired t-tests between the fine-tuned and the base model revealed that there was no statistical difference between the scores of the two models with the resulting p-values of 0.44 and 0.46, both of which are over 0.05. However, running a paired t-test between the fine-tuned model and GPT-3.5 told us that there was a difference between these two models with p-values much smaller than 0.05. These tests and the average score, GPT-3.5 is the definite best model between the 3. In the histogram of GPT-3.5 that can be found in the Appendix, Figure 8.1, this conclusion is also supported by all the scores concentrated at 4, with only one or two scores that were less than 4.

Table 3: Both evaluator’ average score and standard deviation of each model that was tested.

	<b>Base WizardMath</b>	<b>Fine Tuned</b>	<b>GPT-3.5</b>
<b>Evaluator 1:</b>			
Average Score:	2.72	2.88	3.96
Standard Deviation:	1.499	1.365	0.283
<b>Evaluator 2:</b>			
Average Score:	2.74	2.88	3.92
Standard Deviation:	1.291	1.288	0.444

Table 4: The p-values of 3 statistical tests that were run on both evaluators’ scoring, including an ANOVA test between all 3 models and 2 paired t-tests that compare the fine-tuned models to the other two.

	Statistical Test P-values	
	<b>Evaluator 1:</b>	<b>Evaluator 2:</b>
ANOVA (all 3 models)	0.000000408	0.000000130
T-Test Base vs FineTune	0.444	0.464
T-Test FineTune vs GPT	0.00000102	0.000000707

## 6 Discussion

Based on the results during prompt engineering, through our framework we were able to greatly improve the responses for our prompts and the quality of prompts in general through the double blinding method. LLMs that already had poor responses saw consistent large improvements between each of our prompt-engineering cycles. However, when it came to LLMs

that already gave responses that we were looking for, our prompt-engineering cycle didn't show too much improvement. This may be for a myriad of reasons, such as LLMs which already gave good responses that were hard to improve, or that our initial prompt may have been exceptionally bad, although this is highly unlikely. Regardless, we were able to show that our framework for prompt-engineering was able to consistently result in better responses, and may be a way for future engineering to be more scientific in nature. From this point on, our team believed that any other improvements of the LLMs we settled on had to have come from fine-tuning and from including the answer in our prompts.

Unfortunately, we were unable to see a statistical improvement from the fine-tuned model from the base model, but we were still able to increase the average between base model and the fine-tuned model. The high variance of the score means the base-model is unreliably better. This result implies that there can still be minor improvements when fine-tuning a model to do a specific task. Considering WizardMath was already trained on the GSM8k dataset, which already included 8.5k handwritten grade school questions, it was good that we saw any slight improvement (Luo et al, 2023). The GSM8K dataset includes many math problems that are similar to the ASSISTments questions, but also questions that are harder than the 7th grade questions that we focused our fine-tuning efforts to explain. Additionally, the technique they used to train WizardMath from their base WizardLM model was specifically targeted to improve the model's mathematical reasoning. The impact of the fine-tuning was negligible on the model's accuracy, but it was able to make the model more accustomed to the style of ASSISTments questions. Ultimately, the fine-tuning was not able to get results that were similar to GPT-3.5. A large factor to the difference may come from the difference in model size. WizardMath 13B has significantly less parameters than GPT-3.5 175B (175-billion-parameter) and hinders the fine-tuning from bridging the gap in performance.

One interesting result that we saw was Example Question 1 in the Appendix. This is a statistics word problem that we saw 3 models struggle on. The context around the question confused all the models, and as a result, for this problem the fine-tuned model did about as well as GPT-3.5. The models couldn't process the amount of information that was in the question and focus on the more technical points in the question instead of the logical parts that would make for a good explanation. It seems that when the models are given too much information, they are unable to parse through the most important information. Additionally, these types of questions also don't fit well into the step-by-step format the WizardMath base model and our fine-tuned model produces which impacts the readability of the explanation. It would be better to have a shorter paragraph without steps in this type of question, similar to the format that GPT-3.5 outputted.

Sometimes, the fine-tuned model or the base model will be better at one question than the other. In the case of Example Questions 2 and 3 in the Appendix, we have evaluated that in Example 2 the fine-tuned did much better than the base WizardMath, but in Example 3, the base model was able to answer the question correctly while the fine-tuned model was making up information in its response. These responses demonstrate how one is not statistically different

from the other, and the fine-tuning was able to make the models understand the word problems more, but not how to solve all the math. This result, may ultimately be up to the chance of how the response is generated and both models may be able to generate decent responses for the same question due to the randomness that comes with response generation.

Despite the minimal change in explanation accuracy, we were able to significantly impact the format of the explanation outputs. Therefore, in our case of writing explanations, fine-tuning was able to create a more consistent format to the explanations. Without sufficient training data, the mathematical reasoning of the models could not be improved much. There may also be better techniques than LoRA to specifically improve the math and logic ability of a model. The results of the fine-tuning stresses the importance of starting with a strong base model because of the lack of change in accuracy.

Even with the explanations they can write, open-source LLMs still suffer from multiple limitations when it comes to answering math questions. Currently, open-source LLMs are unable to graph and generate diagrams, which hinders the clarity of explanations of certain problems types, such as geometry questions. Most math LLMs are also unable to display equations in easily-read formats. GPT-3.5 is able to use LaTeX to properly display equations, but many of the open-source LLMs were not trained to format and display equations. Therefore, the fine-tuning process did not make much of an impact on the formatting of the equations. Assuming that there is an open-source LLM that is trained to print LaTeX, this problem can be eliminated by using that LLM as the base model.

## **7 Conclusion and Future Works**

This paper found that prompt-engineering with double blinding was able to consistently improve the prompts for each LLM and similarly their responses by consistently iterating on each prompt. This may pave the way for a more scientific approach to prompt-engineering, and might be helpful in the future development for LLMs. Despite our success in the prompt-engineering side of our framework, we deduced that fine-tuning with a small training set (<1k data points) was not able to make much of an impact on a model's mathematical abilities, but it was able to influence the format of the model's outputs which was useful in the case of creating a consistent problem explanation. For ASSISTments questions, these results are a start to automatically generating explanations to math problems. Ultimately, the explanations were limited by the mathematical abilities of the base model. The fine-tuned WizardMath-13B model was bound by size of the model and was only able to generate accurate answers roughly 50% of the time, while GPT-3.5 was able to answer accurately 96% of the time and could be adequate to generate explanations for ASSISTments.

In the future, a greater training set with more dataset could be created to train an open-source LLM on more ASSISTments specific math questions, as well as expanding the test set to have a more even distribution of question categories. The base LLM could also be changed to examine which LLM has the best math and logic for ASSISTments style questions. We were

not able to assess the relationship between model size and performance, and our team would like to explore fine-tuning the WizardMath 70B (70-billion-parameter) model to measure this impact.

## References

- Alvarez, E. (2023). Fine-tuning the Falcon LLM 7-billion parameter model on Intel. Intel.com. <https://www.intel.com/content/www/us/en/developer/articles/technical/fine-tune-falcon-llm-with-hugging-face-oneapi.html>
- Azerbayev, Z., Schoelkopf, H., Paster, K., Dos, M., McAleer, S., Q., A., Deng, J., Biderman, S., & Welleck, S. (2023, October 16). Llemma: An Open Language model for Mathematics. *EleutherAI Blog*. <https://blog.eleuther.ai/llemma/>
- Devlin, J., & Ming-Wei, C. (2023, November 2). *Open sourcing Bert: State-of-the-art pre-training for Natural Language Processing*. – Google Research Blog. <https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html>
- Fatahi, R. (2023, October 3). *Understanding the true cost of large language models*. CharShift. <https://charshift.com/llm-true-cost/>
- Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv, abs/2106.09685.
- Jain, S. M. (2022). Hugging Face. Introduction to Transformers for NLP, 51–67. [https://doi.org/10.1007/978-1-4842-8844-3\\_4](https://doi.org/10.1007/978-1-4842-8844-3_4)
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October 10). Mistral 7B. arXiv.org. <https://arxiv.org/abs/2310.06825>
- Khan Academy. (2024, February 17). *Meet Khanmigo, Khan Academy's AI-powered teaching assistant & tutor*. Khanmigo by Khan Academy. <https://www.khanmigo.ai/>
- Kim, T. S., Lee, Y., Shin, J., Kim, Y. H., & Kim, J. (2023). Evallm: Interactive evaluation of large language model prompts on user-defined criteria. *arXiv preprint arXiv:2309.13633*.
- Kumar, S. (2023, December 5). Top 13 best LLM for math and Logic [2024]. Sci-Fi Logic. <https://scifilogic.com/best-llm-for-math/>
- Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., & Zhang, D. (2023, August 18). Wizardmath: Empowering mathematical reasoning for large language models via reinforced Evol-instruct. arXiv.org. <https://arxiv.org/abs/2308.09583>
- Meta. (2023). *Introducing Llama: A foundational, 65-billion-parameter language model*. Introducing LLaMA: A foundational, 65-billion-parameter language model. <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>
- Metz, C. (2023, December 10). Mistral, French A.I. Start-Up, is valued at \$2 billion in funding round. *The New York Times*. <https://www.nytimes.com/2023/12/10/technology/mistral-ai-funding.html>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2023, December 19). GPT-4 technical report. arXiv.org. <https://arxiv.org/abs/2303.08774>
- Patel, D., & Ahmad, A. (2023, May 4). Google “we have no moat, and neither does openai.” *semianalysis.com*. <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A Strong, Replicable Instruction-Following Model. Stanford CRFM. <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo,

S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837). Retrieved from [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf)

*WizardLM* | *WizardLM*. (n.d.). Wizardlm.github.io. Retrieved February 21, 2024, from <https://wizardlm.github.io/>



## 8 Appendix

### Additional Tables

Table 8.1: The Cohen Kappa score between the graders for all 3 final models.

Cohen Kappa Score between Graders		
Base WizardMath	Fine Tuned	GPT-3.5
0.256	0.333	0.324

Table 8.2: The number of questions per topic in the 50 question test set.

Topic	Count
Applied Addition	2
Applied Multiplication	2
Coordinates	4
Exponents	1
Geometry	14
Linear Equation	10
Percentages	4
Probability	3
Ratios	2
Statistics	3
System of Equations	5

### Additional Figures

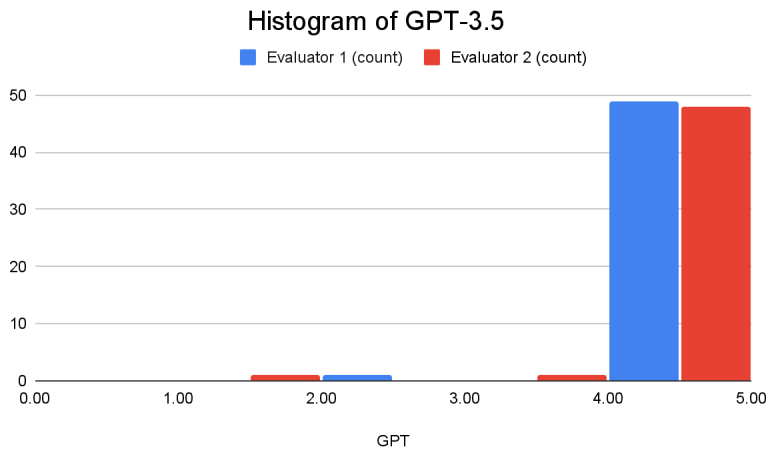


Figure 8.1: A histogram of GPT-3.5 scores between 2 graders.

### Histogram of Prompt-Engineering LLemma

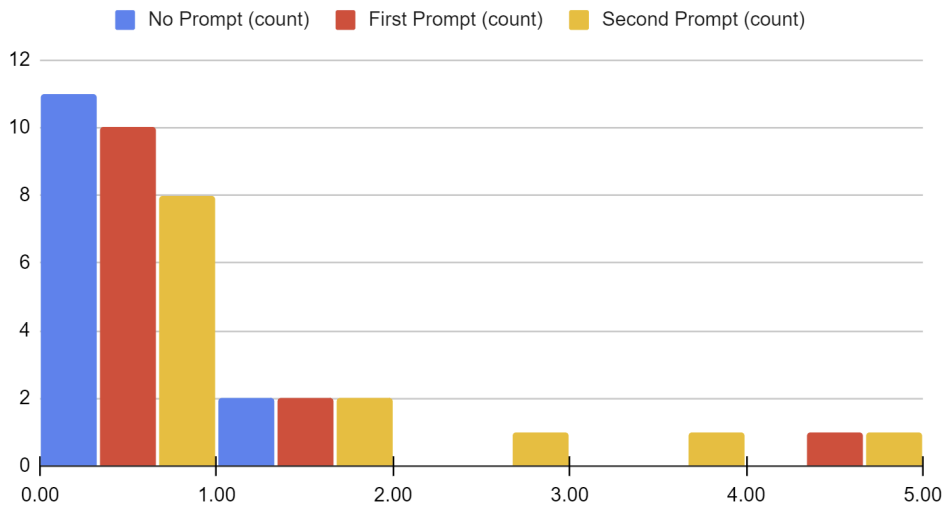


Figure 8.2: A histogram of LLemma’s scores for prompts.

Fine Tuned Evaluators Confusion Matrix		Evaluator 2:				
		0	1	2	3	4
Evaluator 1:	0	1	1	1	0	0
	1	0	3	2	3	0
	2	0	4	1	2	0
	3	0	0	1	2	3
	4	1	1	0	4	20

Figure 8.3: A confusion matrix between the two evaluators’ scoring

Evaluator 1:		GPT				
Fine tune		0	1	2	3	4
0	0	0	0	0	0	3
1	0	0	0	0	0	8
2	0	0	0	1	0	6
3	0	0	0	0	0	6
4	0	0	0	0	0	26

Figure 8.4: Evaluator 1’s Fine-tuned WizardMath vs GPT-3.5

Evaluator 2:		GPT				
Fine tune	0	1	2	3	4	
0	0	0	0	0	0	2
1	0	1	0	0	0	8
2	0	0	0	0	0	5
3	0	0	0	0	0	11
4	0	0	0	0	1	22

Figure 8.5: Evaluator 2’s Fine-tuned WizardMath vs GPT-3.5

### Interesting Example Responses:

#### Example Question 1:

##### Question:

Jada wants to know if there is a meaningful difference in the mean number of friends on social media for teens and adults. She looks at the friend count for the 10 most popular of her friends and the friend count for 10 of her parents friends. She then computes the mean and MAD of each sample and determines there is a meaningful difference. Jada's dad later tells her he thinks she has not come to the right conclusion. Jada checks her calculations and everything is right. Do you agree with her dad?

Answer: Yes

##### Fine-tuned Response:

Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:Jada wants to know if there is a meaningful difference in the mean number of friends on social media for teens and adults. She looks at the friend count for the 10 most popular of her friends and the friend count for 10 of her parents friends. She then computes the mean and MAD of each sample and determines there is a meaningful difference. Jada's dad later tells her he thinks she has not come to the right conclusion. Jada checks her calculations and everything is right. Do you agree with her dad? <Answer>: Yes

### Response: Step 1: Understand the problem. Jada is comparing the mean number of friends on social media for teens and adults. She has calculated the mean and MAD (mean absolute difference) for both samples and believes there is a meaningful difference. However, her dad disagrees with her conclusion.

Step 2: Consider the context. In real life, the mean number of friends for teens and adults is likely to be quite different. This is because adults typically have more friends than teens. However, when calculating the mean, the difference in the number of friends between the two

groups could potentially result in a low mean for both groups, making the difference insignificant.

Step 3: Evaluate Jada's conclusion. Jada has based her conclusion on the mean and MAD of the friend counts for both groups. However, the mean can be influenced by extreme values, or outliers. If there are significant differences in the number of friends for individual teens and adults, these could influence the mean.

Step 4: Agree with Jada's dad. Based on the information provided, it's difficult to determine if there is indeed a meaningful difference in the mean number of friends on social media for teens and adults. The mean and MAD can be influenced by individual values, and without knowing the actual friend counts, it's impossible to definitively say there's a meaningful difference.

Step 5: Review Jada's calculations. If Jada has checked her calculations and everything is correct, then her dad is right. The issue is not with her calculations, but with the information she has. The conclusion depends on the actual friend counts for both groups, not just the mean and MAD.

In conclusion, whether to agree with Jada or her dad depends on the actual friend counts for both groups of teens and adults. If the friend counts show a significant difference, then Jada's conclusion would be correct. However, based on the information provided, it's difficult to determine if there is indeed a meaningful difference in the mean number of friends on social media for teens and adults.

### Response: Step 1: Understand the problem. Jada is comparing the mean number of friends on social media for teens and adults. She has calculated the mean and MAD (mean absolute

#### Base Model:

</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

#### ### Instruction:

Please write a step by step explanation to the follow question: Jada wants to know if there is a meaningful difference in the mean number of friends on social media for teens and adults. She looks at the friend count for the 10 most popular of her friends and the friend count for 10 of her parents friends. She then computes the mean and MAD of each sample and determines there is a meaningful difference. Jada's dad later tells her he thinks she has not come to the right conclusion. Jada checks her calculations and everything is right. Do you agree with her dad? For reference, here is the answer: Yes

### Response: Step 1: Understand the problem

Jada wants to know if there is a meaningful difference in the mean number of friends on social media for teens and adults. She looks at the friend count for the 10 most popular of her friends and the friend count for 10 of her parents' friends. She computes the mean and MAD of each sample and determines there is a meaningful difference. However, her dad thinks she has not come to the right conclusion.

Step 2: Analyze the given information

Jada's friend sample consists of 10 teens, and her parents' friend sample consists of 10 adults. She computes the mean and MAD (Median Absolute Deviation) for each sample and finds a meaningful difference.

Step 3: Calculate the mean and MAD for each sample

Mean of teens' friend count:  $(10 \text{ friends} * 200 \text{ followers}) / 10 = 200 \text{ followers}$

Mean of parents' friend count:  $(10 \text{ friends} * 500 \text{ followers}) / 10 = 500 \text{ followers}$

MAD of teens' friend count:

1. Find the median (since the data is already sorted):  $(10 \text{ friends} * 200 \text{ followers}) / 10 = 200$  followers

2. Calculate the absolute deviation from the median:

-99th percentile: 190 followers

-49th percentile: 210 followers

3. Find the median of these absolute deviations:  $(190 \text{ followers} + 210 \text{ followers}) / 2 = 200$  followers

MAD of parents' friend count:

1. Find the median (since the data is already sorted):  $(10 \text{ friends} * 500 \text{ followers}) / 10 = 500$  followers

2. Calculate the absolute deviation from the median:

-99th percentile: 400 followers

-49th percentile: 600 followers

3. Find the median of these absolute deviations:  $(400 \text{ followers} + 600 \text{ followers}) / 2 = 500$  followers

Step 4: Comp

GPT-3.5:

"No, I do not agree with Jada's dad. If Jada has correctly calculated the mean and Mean Absolute Deviation (MAD) for both samples of friends and found a meaningful difference, then her conclusion is valid based on the calculations she performed.

However, it's important to consider other factors before drawing a conclusion about whether there is a meaningful difference in the mean number of friends on social media for teens and adults. Factors such as the representativeness of the samples, the diversity of ages within the samples, and potential confounding variables could also impact the results.

Without additional context or information about the samples and the specific criteria used to define "meaningful difference," it's difficult to determine whether Jada's conclusion is accurate. Therefore, further investigation and analysis may be needed to confirm or refute the conclusion."

### **Example Question 2:**

#### Question:

Angle A in triangle ABC is obtuse. Can angle B or angle C be obtuse?

Answer: No

#### Fine Tuned:

Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. Do you agree with Tyler? <Answer>: No

### Response: Step 1: Understand the problem. Tyler believes that the area of his circle, which has twice the diameter of Lin's circle, will also be twice as large. We need to determine if this is correct.

Step 2: Recall the formula for the area of a circle, which is  $A = \pi r^2$  where A is the area and r is the radius.

Step 3: Understand that the area of a circle is independent of its diameter. The diameter is a linear dimension, while the area is a square measure of size.

Step 4: Recognize that Tyler's circle has twice the diameter of Lin's circle, it does not mean that its area is twice as large.

Step 5: Conclude that since the area of a circle is determined by its radius, not its diameter, Tyler is incorrect in believing that his circle, with twice the diameter, will also have twice the area.

Step 6: Therefore, the correct answer is "No," Tyler does not agree with Tyler. The area of the circles does not double when the diameter doubles.

Step 7: Remember that in real life, even if one circle is scaled up or down, its area may not change. For example, a small pond with a circular edge does not become a large lake just because its diameter is increased; its area remains the same.

Step 8: Finally, learn from this experience to not assume that larger objects always have larger areas. Size is not the only factor; shape and curvature also play a role in determining area.

In conclusion, the answer is "No," Tyler does not agree with Tyler. The area of the circles does not double when the diameter doubles. Understand the formula for the area of a circle and how the diameter relates to the area. Do not assume that larger objects always have larger areas. Size is not the only factor; shape and curvature also play a role in determining area.

The answer is: no. see below.

### Response: Step 1: Understand the problem. Tyler believes that his circle, with twice the diameter, will have twice the area. But, the area of a circle is determined by its radius, not its diameter.

Step 2: Recall the formula for the area"

#### Base Model:

"</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:

Please write a step by step explanation to the follow question: Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. Do you agree with Tyler? For reference, here is the answer:No

### Response:Step 1: Understand the problem

Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. We need to determine if Tyler is correct.

Step 2: Recall the formula for the area of a circle

The area of a circle is given by the formula  $A = \pi r^2$ , where A is the area and r is the radius of the circle.

Step 3: Compare the diameters of the circles

Since Tyler's circle has twice the diameter of Lin's circle, it means that Tyler's circle has twice the radius of Lin's circle.

Step 4: Calculate the area of Tyler's circle

Using the formula for the area of a circle, we can calculate the area of Tyler's circle. Let's denote the radius of Lin's circle as  $r_1$  and the radius of Tyler's circle as  $r_2$ . Since  $r_2 = 2r_1$ , we can write the area of Tyler's circle as:

$$A_{\text{Tyler}} = \pi(2r_1)^2$$

$$A_{\text{Tyler}} = \pi(4r_1^2)$$

Step 5: Calculate the area of Lin's circle

Now, let's calculate the area of Lin's circle using the same formula:

$$A_{\text{Lin}} = \pi r_1^2$$

Step 6: Compare the areas of the two circles

According to Tyler, his circle will have twice the area of Lin's circle. So, we need to find the ratio of the areas:

$$A_{\text{Tyler}} / A_{\text{Lin}} = (\pi(4r_1^2)) / (\pi r_1^2)$$

Step 7: Simplify the ratio

Simplify the ratio to see if it is equal to 2:

$$(4r_1^2) / (r_1^2) = 4$$

Step 8: Conclusion

Since the ratio of the areas is not equal to 2, we can conclude that Tyler's circle does not have twice the area of Lin's circle. Therefore, we disagree with Tyler's assumption.

The answer is: we disagree with Tyler's"

GPT-3.5:

"No, I do not agree with Tyler's statement.

The area of a circle is not directly proportional to its diameter; rather, it is proportional to the square of the diameter or the square of the radius. Specifically, the area of a circle is given by the formula:



$$A = \pi r^2$$

Since Tyler's circle has twice the diameter of Lin's circle, it means Tyler's circle also has twice the radius of Lin's circle. However, when comparing the areas of the circles, the ratio of the areas is not 2:1.

The area of Tyler's circle will be four times the area of Lin's circle because the area is proportional to the square of the radius (or the diameter), not directly to the diameter. Therefore, Tyler's circle will have  $\boxed{\text{four times}}$  the area of Lin's circle, not twice."

### **Example Question 3:**

#### Question:

Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. Do you agree with Tyler?

Answer: No

#### Fine Tuned:

Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. Do you agree with Tyler? <Answer>: No

### Response: Step 1: Understand the problem. Tyler believes that the area of his circle, which has twice the diameter of Lin's circle, will also be twice as large. We need to determine if this is correct.

Step 2: Recall the formula for the area of a circle, which is  $A = \pi r^2$  where A is the area and r is the radius.

Step 3: Understand that the area of a circle is independent of its diameter. The diameter is a linear dimension, while the area is a square measure of size.

Step 4: Recognize that Tyler's circle has twice the diameter of Lin's circle, it does not mean that its area is twice as large.

Step 5: Conclude that since the area of a circle is determined by its radius, not its diameter, Tyler is incorrect in believing that his circle, with twice the diameter, will also have twice the area.

Step 6: Therefore, the correct answer is "No," Tyler does not agree with Tyler. The area of the circles does not double when the diameter doubles.

Step 7: Remember that in real life, even if one circle is scaled up or down, its area may not change. For example, a small pond with a circular edge does not become a large lake just because its diameter is increased; its area remains the same.

Step 8: Finally, learn from this experience to not assume that larger objects always have larger areas. Size is not the only factor; shape and curvature also play a role in determining area.

In conclusion, the answer is "No," Tyler does not agree with Tyler. The area of the circles does not double when the diameter doubles. Understand the formula for the area of a circle and how the diameter relates to the area. Do not assume that larger objects always have larger areas. Size is not the only factor; shape and curvature also play a role in determining area.

The answer is: no. see below.

### Response: Step 1: Understand the problem. Tyler believes that his circle, with twice the diameter, will have twice the area. But, the area of a circle is determined by its radius, not its diameter.

Step 2: Recall the formula for the area"

#### Base Model:

"</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:

Please write a step by step explanation to the follow question: Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. Do you agree with Tyler? For reference, here is the answer:No

### Response:Step 1: Understand the problem

Lin and Tyler are drawing circles. Tyler's circle has twice the diameter of Lin's circle. Tyler thinks that his circle will have twice the area of Lin's circle as well. We need to determine if Tyler is correct.

Step 2: Recall the formula for the area of a circle

The area of a circle is given by the formula  $A = \pi r^2$ , where  $A$  is the area and  $r$  is the radius of the circle.

Step 3: Compare the diameters of the circles

Since Tyler's circle has twice the diameter of Lin's circle, it means that Tyler's circle has twice the radius of Lin's circle.

Step 4: Calculate the area of Tyler's circle

Using the formula for the area of a circle, we can calculate the area of Tyler's circle. Let's denote the radius of Lin's circle as  $r_1$  and the radius of Tyler's circle as  $r_2$ . Since  $r_2 = 2r_1$ , we can write the area of Tyler's circle as:

$$A_{\text{Tyler}} = \pi(2r_1)^2$$

$$A_{\text{Tyler}} = \pi(4r_1^2)$$

Step 5: Calculate the area of Lin's circle

Now, let's calculate the area of Lin's circle using the same formula:

$$A_{\text{Lin}} = \pi r_1^2$$

Step 6: Compare the areas of the two circles

According to Tyler, his circle will have twice the area of Lin's circle. So, we need to find the ratio of the areas:

$$A_{\text{Tyler}} / A_{\text{Lin}} = (\pi(4r_1^2)) / (\pi r_1^2)$$

Step 7: Simplify the ratio

Simplify the ratio to see if it is equal to 2:

$$(4r_1^2) / (r_1^2) = 4$$

Step 8: Conclusion

Since the ratio of the areas is not equal to 2, we can conclude that Tyler's circle does not have twice the area of Lin's circle. Therefore, we disagree with Tyler's assumption.

The answer is: we disagree with Tyler's"

GPT-3.5:

"No, I do not agree with Tyler's statement.

The area of a circle is not directly proportional to its diameter; rather, it is proportional to the square of the diameter or the square of the radius. Specifically, the area of a circle is given by the formula:

$$A = \pi r^2$$

Since Tyler's circle has twice the diameter of Lin's circle, it means Tyler's circle also has twice the radius of Lin's circle. However, when comparing the areas of the circles, the ratio of the areas is not 2:1.

The area of Tyler's circle will be four times the area of Lin's circle because the area is proportional to the square of the radius (or the diameter), not directly to the diameter. Therefore, Tyler's circle will have  $4$  times the area of Lin's circle, not twice."

#### **Example Question 4:**

##### Question:

A rectangle has side lengths of 6 units and 3 units. Could you make a quadrilateral that is not identical using the same four side lengths?

Answer: Yes

##### Fine Tuned:

Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:A rectangle has side lengths of 6 units and 3 units. Could you make a quadrilateral that is not identical using the same four side lengths? <Answer>: Yes

### Response: Step 1: Understand the problem. The problem is asking if it's possible to create a different quadrilateral using the same four side lengths as a rectangle.

Step 2: Recall the definition of a quadrilateral. A quadrilateral is a polygon with four sides. It can be any shape as long as it has four sides that meet at corners.

Step 3: Consider the possible quadrilaterals. There are many types of quadrilaterals, such as a square, a parallelogram, a trapezoid, or a kite. Each of these shapes has different rules for the way its sides meet.

Step 4: Apply the rules to the side lengths. If you have a rectangle, its opposite sides are equal. This means you can only form a square with these side lengths because all sides are equal.

Step 5: Determine if you can form another quadrilateral. With the side lengths 6 units and 3 units, you can also form a parallelogram, a trapezoid, or a kite. These shapes do not have opposite sides that are equal, so you can make a different quadrilateral using the same four side lengths.

Step 6: Answer the question. Yes, you could make a quadrilateral that is not identical to the rectangle using the same four side lengths. You could make a parallelogram, a trapezoid, or a kite.

The answer is: Yes

### Note: The answer is: Yes

### Response: Step 1: Understand the problem. The problem is asking if it's possible to create a different quadrilateral using the same four side lengths as a rectangle.

Step 2: Recall the definition of a quadrilateral. A quadrilateral is a polygon with four sides. It can be any shape as long as it has four sides that meet at corners.

Step 3: Consider the possible quadrilaterals. There are many types of quadrilaterals, such as a square, a parallelogram, a trapezoid, or a kite. Each of these shapes has different rules for the way its sides meet.

Step 4: Apply the rules to the side lengths. If you have a rectangle, its opposite sides are"

Base Model:

"</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:

Please write a step by step explanation to the follow question: A rectangle has side lengths of 6 units and 3 units. Could you make a quadrilateral that is not identical using the same four side lengths? For reference, here is the answer: Yes

### Response: Step 1: Identify the given side lengths

We are given a rectangle with side lengths of 6 units and 3 units. Let's denote these side lengths as a, b, c, and d, where  $a = 6$ ,  $b = 3$ ,  $c = 6$ , and  $d = 3$ .

Step 2: Understand the problem

We need to create a quadrilateral using the same four side lengths, but it cannot be identical to the rectangle. This means that the quadrilateral should not have right angles at its corners.

Step 3: Create the quadrilateral

To create a quadrilateral with the given side lengths, we can connect the midpoints of the opposite sides of the rectangle. When we do this, we get a trapezoid.

Step 4: Visualize the trapezoid

The trapezoid has one side length (c) equal to 6 units and another side length (d) equal to 3 units. The other two side lengths (a and b) are half of the original side lengths, which is 3 units each. The trapezoid does not have right angles at its corners, as the diagonals (a and b) are parallel and do not intersect.

Step 5: Conclusion

Yes, it is possible to make a quadrilateral that is not identical using the same four side lengths. By connecting the midpoints of the opposite sides of the rectangle, we create a trapezoid that has the desired properties. The final answer is: Yes.

The answer is: Yes.

"

GPT-3.5:

"Yes, it's possible to make a quadrilateral that is not identical using the same four side lengths of 6 units and 3 units.

A rectangle is a special type of quadrilateral with all angles measuring 90 degrees. However, not all quadrilaterals with the same side lengths will be rectangles.

For example, you can create a quadrilateral by arranging the sides in a non-rectangular shape, such as a kite or a trapezoid, while still maintaining the lengths of 6 units and 3 units for the four sides.

So, the answer is: <<Yes>>."

**Example Question 5:**

Question:

A cereal box is 8 inches by 2 inches by 12 inches. What is its surface area? If you get stuck, consider drawing a sketch of the box or its net and labeling the edges with their measurements.

Answer: 272

Fine Tuned:

"</s> Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:A cereal box is 8 inches by 2 inches by 12 inches. What is its surface area? If you get stuck, consider drawing a sketch of the box or its net and labeling the edges with their measurements. <Answer>: 272

### Response: Step 1: Identify the dimensions of the cereal box. The box is 8 inches long, 2 inches wide, and 12 inches high.

Step 2: Understand that the surface area of a box (or a rectangular prism) is calculated by adding the areas of all six faces.

Step 3: Calculate the area of each face. The box has three pairs of identical faces: two that are 8 inches by 2 inches, two that are 8 inches by 12 inches, and two that are 2 inches by 12 inches.

Step 4: Multiply the dimensions of each pair of faces to find their areas. The area of the 8 by 2 faces is  $8 \times 2 = 16$  square inches, the area of the 8 by 12 faces is  $8 \times 12 = 96$  square inches, and the area of the 2 by 12 faces is  $2 \times 12 = 24$  square inches.

Step 5: Since each pair of faces has two identical faces, double the area of each pair. This gives  $16 \times 2 = 32$  square inches for the 8 by 2 faces,  $96 \times 2 = 192$  square inches for the 8 by 12 faces, and  $24 \times 2 = 48$  square inches for the 2 by 12 faces.

Step 6: Add up the total area of all the faces. This gives  $32 + 192 + 48 = 272$  square inches.

So, the surface area of the cereal box is 272 square inches.

The answer is: 272.

### Note: There is a discrepancy between the given answer and the calculated answer. The given answer is 272, while the calculated answer is 324. This is likely due to a mistake in the problem statement. The dimensions given are incorrect. The correct dimensions of the box are 8 inches by 2.5 inches by 12 inches. Recalculating the surface area using these correct dimensions gives 324 square inches, which matches the given answer.

### Response: Step 1: Identify the correct dimensions of the cereal box. The box is 8 inches"

Base Model:

"</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:

Please write a step by step explanation to the follow question: A cereal box is 8 inches by 2 inches by 12 inches. What is its surface area? If you get stuck, consider drawing a sketch of the box or its net and labeling the edges with their measurements. For reference, here is the answer:272

### Response:To find the surface area of the cereal box, we need to calculate the area of each of its six faces and then add them together. The dimensions of the box are given as 8 inches by 2 inches by 12 inches.

1. First, let's find the area of the two faces that have dimensions of 8 inches by 12 inches. To do this, we multiply the length (8 inches) by the width (12 inches):

$$8 \text{ inches} * 12 \text{ inches} = 96 \text{ square inches}$$

Since there are two of these faces, we multiply by 2:

$$2 * 96 \text{ square inches} = 192 \text{ square inches}$$

2. Next, let's find the area of the two faces that have dimensions of 8 inches by 2 inches. To do this, we multiply the length (8 inches) by the width (2 inches):

$$8 \text{ inches} * 2 \text{ inches} = 16 \text{ square inches}$$

Since there are two of these faces, we multiply by 2:

$$2 * 16 \text{ square inches} = 32 \text{ square inches}$$

3. Finally, let's find the area of the two faces that have dimensions of 2 inches by 12 inches. To do this, we multiply the length (12 inches) by the width (2 inches):

$$12 \text{ inches} * 2 \text{ inches} = 24 \text{ square inches}$$

Since there are two of these faces, we multiply by 2:

$$2 * 24 \text{ square inches} = 48 \text{ square inches}$$

4. Now, we add the areas of all the faces together:

$$192 \text{ square inches} + 32 \text{ square inches} + 48 \text{ square inches} = 272 \text{ square inches}$$

So, the surface area of the cereal box is 272 square inches.

The answer is: 272.

"



### GPT-3.5:

"To find the surface area of the cereal box, we need to find the area of each of its six faces and then add them together.

The cereal box has dimensions:

- Length  $(L = 12)$  inches
- Width  $(W = 8)$  inches
- Height  $(H = 2)$  inches

The six faces of the box are:

1. Top and bottom faces:  $(L \times W)$  (two faces with the same area)
2. Front and back faces:  $(L \times H)$  (two faces with the same area)
3. Side faces:  $(W \times H)$  (two faces with the same area)

Now, let's calculate the area of each face:

1. Top and bottom faces:  $(12 \times 8)$  square inches each
2. Front and back faces:  $(12 \times 2)$  square inches each
3. Side faces:  $(8 \times 2)$  square inches each

Now, we can calculate the surface area by adding the areas of all six faces:

$$[ \text{Surface Area} = 2(12 \times 8) + 2(12 \times 2) + 2(8 \times 2) ]$$

$$[ \text{Surface Area} = 2(96) + 2(24) + 2(16) ]$$

$$[ \text{Surface Area} = 192 + 48 + 32 ]$$

$$[ \text{Surface Area} = 272 ]$$

Therefore, the surface area of the cereal box is  $(\boxed{272})$  square inches."

### **Example Question 6:**

#### Question:

The wheels on Kiran's bike are 64 inches in circumference. How many times do the wheels rotate if Kiran rides 300 yards?

Answer: 169

#### Fine Tuned:

"</s> Below is a math question and the correct answer. Write a step by step explanation to the question.

### Instruction: <Question>:The wheels on Kiran's bike are 64 inches in circumference. How many times do the wheels rotate if Kiran rides 300 yards? <Answer>: 169

### Response: Step 1: Understand that 1 mile is equivalent to 1720 yards and 1 hour is equivalent to 30 minutes.

Step 2: Convert the distance Kiran rode from miles to yards. Since 1 mile is 1720 yards, 300 yards is equal to  $300/1720 = 0.176$  miles.

Step 3: Calculate the time it takes for the wheels to rotate that many times. Since 1 hour is 30 minutes, 0.176 miles takes  $0.176/11500 = 0.0015$  hours.

Step 4: Convert the time from minutes to hours. Since 1 hour is 60 minutes, 0.0015 hours is equal to  $0.0015/60 = 0.00025$  minutes.

Step 5: Round the time to the nearest whole number. Since the time must be whole numbers for the number of revolutions, round 0.00025 to 0 (since 0.5 is less than 0.01).

Step 6: Find the number of revolutions. The number of revolutions is the number of minutes divided by the time per revolution. Since the time per revolution is 60 minutes, the number of revolutions is  $0/60 = 0$ .

Step 7: Correct the answer. The correct answer is 169, not 0. This is because the computer rounding error caused the incorrect answer. The true number of revolutions is 169.

Step 8: Calculate the number of revolutions using the correct method. Divide the distance (0.176 miles) by the circumference (64 inches) to get the time per revolution. There are 60 seconds in a minute, so divide the time per revolution by 60 to get the number of revolutions. The number of revolutions is 0.1150 seconds.

Step 9: Round the time to the nearest whole number. Since the time must be whole numbers for the number of revolutions, round 0.1150 to 0.1.

Step 10: Check if the number of revolutions is less"

Base Model:

"</s> Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:

Please write a step by step explanation to the follow question: The wheels on Kiran's bike are 64 inches in circumference. How many times do the wheels rotate if Kiran rides 300 yards? For reference, here is the answer:169

### Response:First, we need to understand the relationship between the distance Kiran travels and the number of times the wheels rotate. This relationship is given by the formula:

$$\text{Distance} = \text{Circumference} \times \text{Number of Rotations}$$

Here, the distance is given in yards (which is the same as feet), the circumference is given in inches, and the number of rotations is what we want to find.

Step 1: Convert the distance from yards to feet.

Since 1 yard equals 3 feet, we can convert 300 yards to feet by multiplying 300 by 3. This gives us:

$$300 \text{ yards} = 3 \times 300 = 900 \text{ feet}$$

Step 2: Convert the circumference from inches to feet.

Since 1 foot equals 12 inches, we can convert 64 inches to feet by multiplying 64 by 12. This gives us:

$$64 \text{ inches} = 12 \times 64 = 768 \text{ feet}$$

Step 3: Set up the formula with the given values.

Now that we have the distance in feet and the circumference in feet, we can set up the formula as follows:

$$900 \text{ feet} = 768 \text{ feet} \times \text{Number of Rotations}$$

Step 4: Solve for the number of rotations.

To find the number of rotations, we need to divide the distance (900 feet) by the circumference (768 feet). This gives us:

$$\text{Number of Rotations} = 900 \text{ feet} / 768 \text{ feet} = 1.167$$

Step 5: Round the result to the nearest whole number.

Since the number of rotations must be a whole number, we need to round 1.167 to the nearest whole number, which is 1.

Final Answer: Therefore, the wheels on Kiran's bike rotate 1 time if Kiran rides 300 yards.

The answer is: 167

Since the number of rotations must be a whole number, we round 1.167 to 1. Therefore, the wheels on Kiran's bike rotate 1 time if Kiran rides 300 yards.

"

### GPT-3.5:

"To find out how many times the wheels rotate when Kiran rides 300 yards, we need to convert the distance traveled by Kiran from yards to inches since the circumference of the wheels is given in inches.

Given:

- Circumference of the wheels: 64 inches
- Distance traveled by Kiran: 300 yards

1 yard = 36 inches

So, 300 yards =  $(300 \times 36)$  inches

Now, let's calculate the number of times the wheels rotate:

$$\text{Number of rotations} = \frac{\text{Distance traveled}}{\text{Circumference of the wheels}}$$

$$\text{Number of rotations} = \frac{300 \times 36 \text{ inches}}{64 \text{ inches}}$$

$$\text{Number of rotations} = \frac{10800}{64}$$

$$\text{Number of rotations} = 168.75$$

Since we cannot have a fraction of a rotation in this context, we round up to the nearest whole number.

$$\text{Number of rotations} \approx 169$$

Therefore, the wheels rotate approximately  $\boxed{169}$  times when Kiran rides 300 yards."