MEASURING MATHEMATICAL PROBLEM SOLVING WITH THE MATH DATASET

Dan Hendrycks UC Berkeley
Collin Burns UC Berkeley
Saurav Kadavath UC Berkeley
Akul Arora UC Berkeley
Steven Basart UChicago

Eric Tang UC Berkeley
Dawn Song UC Berkeley
Jacob Steinhardt UC Berkeley

ABSTRACT

Many intellectual endeavors require mathematical problem solving, but this skill remains beyond the capabilities of computers. To measure this ability in machine learning models, we introduce MATH, a new dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations. To facilitate future research and increase accuracy on MATH, we also contribute a large auxiliary pretraining dataset which helps teach models the fundamentals of mathematics. Even though we are able to increase accuracy on MATH, our results show that accuracy remains relatively low, even with enormous Transformer models. Moreover, we find that simply increasing budgets and model parameter counts will be impractical for achieving strong mathematical reasoning if scaling trends continue. While scaling Transformers is automatically solving most other text-based tasks, scaling is not currently solving MATH. To have more traction on mathematical problem solving we will likely need new algorithmic advancements from the broader research community. The full paper is available at https://arxiv.org/abs/2103.03874.

1 INTRODUCTION

Mathematics is a highly effective tool in many intellectual endeavors. It enables us to count and quantify objects, and it can be relied upon because it is consistent and based on logic. Mathematics pervades the sciences and can be used to model planetary orbits, atomic motion, signal frequencies, and much more. These phenomena can be encoded with mathematics precisely and concisely. This has even led some to describe mathematics as being "unreasonably effective" (Wigner, 1960). These observations speak to the broad reach and domain-generality of mathematics.

In machine learning, mathematics is a valuable testbed for *problem-solving ability*: the ability to analyze a problem, pick out good heuristics from a large set of possibilities, and chain them together to produce an answer. This contrasts with plug-and-chug calculations, a skill which ML models can already exhibit (Henighan et al., 2020). Visual or linguistic reasoning may involve limited problem-solving ability for tasks such as image classification, but unlike math this is not the focus of these domains.

To measure the problem-solving ability of machine learning models, we introduce the MATH dataset, which consists of $12,500$ problems from high school math competitions. Given a problem from MATH, machine learning models generate a sequence, such as `$\frac{2}{3}$`, that encodes the final answer. These answers are unique after normalization, allowing MATH to be scored with exact match rather than with heuristic metrics such as BLEU. In addition, MATH problems are tagged by difficulty from 1 to 5, and span seven subjects including geometry, where diagrams can be specified in text with the Asymptote language. This enables a fine-grained assessment of mathematical problem-solving ability across difficulties and subjects. Finally, problems come with full step-by-step solutions, which are a valuable additional source of training data.

Metamath Theorem Proving

$n \in \mathbb{N} \wedge \frac{n+1}{2} \in \mathbb{N} \implies \exists m \in \mathbb{N} : n = 2m + 1.$
GPT-f's generated proof:

```
|- ((N e. NN0 /\ ((N + 1)/2)  e.
   NN0) -> ((N - 1) / 2) e. NN0)
|- (N e. NN0 -> N e. CC)
|- 1 e. CC
|- ((N e. CC /\ 1 e. CC) ->
   (N - 1) e. CC )
                ⋮
```

DeepMind Mathematics Dataset

```
Divide 1136975704 by -142121963
Answer: -8
Calculate ((-2)/3)/(-1-(-24)/9)
Answer: -2/5
Let k(u) = u**2+u-4. Find k(0)
Answer: -4
Sort 2, 4, 0, 6
Answer: 0, 2, 4, 6
```

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?
Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.
Problem: If $\sum_{n=0}^{\infty} \cos^{2n} \theta = 5$, what is $\cos 2\theta$?
Solution: Note $1 + \cos^2 \theta + \cos^4 \theta + \cdots = \frac{1}{1-\cos^2 \theta} = 5$. Hence, $\cos^2 \theta = \frac{4}{5}$. Then $\cos 2\theta = 2\cos^2 \theta - 1 = \boxed{\frac{3}{5}}$.
Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.
Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then
$\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) =$
$1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

Figure 1: Previous work is based on formal theorem provers or straightforward plug-and-chug problems. Our dataset, MATH, has competition mathematics problems with step-by-step solutions written in LaTeX and natural language.

The MATH dataset is challenging: large language models achieved accuracies ranging from 2.9% to 6.9%. Despite these low accuracies, models clearly possess some mathematical knowledge: they achieve up to 15% accuracy on the easiest difficulty level, and they are able to generate step-by-step solutions that are coherent and on-topic even when incorrect. A computer science PhD attained approximately 40% on MATH, and a three-time IMO gold medalist attained 90%, showing that MATH can be challenging for humans and machines.

The presence of step-by-step solutions allows models to utilize "scratch space": rather than having to generate a final answer immediately, models can first generate solutions that may contain intermediate computations. Interestingly, we found that generating solutions actually *decreased* accuracy relative to immediately outputting a final answer. In contrast, using solutions at training time increases relative accuracy by 10%. Models also do better with hints that contain a prefix of the solution. This shows that models understand and make use of step-by-step solutions, but are unable to wield them of their own accord. Bridging this gap poses an interesting direction for further research.

While MATH covers advanced problem-solving techniques, models arguably also need to be trained thoroughly on the fundamentals of mathematics. To address this, we create the first large-scale mathematics pretraining dataset with hundreds of thousands of step-by-step solutions in natural language and LaTeX. We call this dataset the Auxiliary Mathematics Problems and Solutions (AMPS) pretraining corpus, which consists of Khan Academy and Mathematica data. AMPS has over $100,000$ Khan Academy problems with step-by-step solutions in LaTeX; these exercises are used to teach human students concepts ranging from basic addition to Stokes' Theorem. It also contains over 5 million problems generated using Mathematica scripts, based on 100 hand-designed modules covering topics such as conic sections, div grad and curl, KL divergence, eigenvalues, polyhedra, and Diophantine equations. In total AMPS contains 23GB of problems and solutions. Domain-specific pretraining (Gururangan et al., 2020) on AMPS improves relative accuracy by around 25%, equivalent to a $15\times$ increase in model size.

Altogether, while large Transformer models (Vaswani et al., 2017) make some progress on the MATH dataset, such as by AMPS pretraining or by training with step-by-step solutions, accuracy nonetheless remains relatively low. While enormous Transformers pretrained on massive datasets can now solve most existing text-based tasks, this low accuracy indicates that our MATH dataset is distinctly harder. Accuracy also increases only modestly with model size: assuming a log-linear scaling trend, models

Model	Prealgebra	Algebra	Number Theory	Counting & Probability	Geometry	Intermediate Algebra	Precalculus	Average
GPT-2 (0.1B)	5.2	5.1	5.0	2.8	5.7	6.5	7.3	5.4 (+0%)
GPT-2 (0.3B)	6.7	6.6	5.5	3.8	6.9	6.0	7.1	6.2 (+15%)
GPT-2 (0.7B)	6.9	6.1	5.5	5.1	8.2	5.8	7.7	6.4 (+19%)
GPT-2 (1.5B)	8.3	6.2	4.8	5.4	8.7	6.1	8.8	6.9 (+28%)
GPT-3 (2.7B)	2.8	2.9	3.9	3.6	2.1	2.5	2.6	2.9 (−46%)
GPT-3 (175B)	7.7	6.0	4.4	4.7	3.1	4.4	4.0	5.2 (−4%)

Table 1: MATH accuracies across subjects for GPT-2 and *few-shot* GPT-3 models. The character 'B' denotes the number of parameters in billions. The gray text indicates the *relative* improvement over the 0.1B baseline. All GPT-2 models pretrain on AMPS, and all values are percentages. A $15\times$ increase in model parameters increased accuracy by 1.5%, a 28% relative improvement. Likewise, enormous GPT-3 models do not automatically solve the MATH benchmark, unlike many other benchmarks. Model accuracy is growing slowly and is far from the ceiling, so much future research is needed.

would need around 10^{35} parameters to achieve 40% accuracy on math, which is impractical. Instead, to make large strides on the MATH dataset with a practical amount of resources, we will need new algorithmic advancements from the broader research community.

2 The MATH Dataset

In this section, we introduce two new datasets, one for pretraining (AMPS) and one for benchmarking mathematical problem-solving ability (MATH). The problems in AMPS can help teach models plug-and-chug calculations. This is a prerequisite for MATH, which goes beyond plug-and-chug questions to test mathematical problem-solving ability.

The MATH dataset consists of problems from mathematics competitions including the AMC 10, AMC 12, AIME, and more. Many of these competition problems can be collected from artofproblemsolving.com/community/c3158_usa_contests. The competitions span decades and assess the mathematical problem-solving ability of the best mathematical talent in the United States. Unlike most prior work, most problems in MATH cannot be solved with a straightforward application of standard K-12 mathematics tools. Instead, humans often solve such problem by applying problem solving techniques and "heuristics" (Pólya, 1945).

The Mathematics Aptitude Test of Heuristics dataset, abbreviated MATH, has $12,500$ problems and step-by-step solutions ($7,500$ training + $5,000$ test). With this many training problems, models can learn many useful heuristics for problem solving. Each problem has a step-by-step solution and a final boxed answer. Example problems with step-by-step solutions are shown in Figure 1.

We further describe the MATH dataset and our AMPS pretraining dataset in the appendix.

3 Experiments

In this section, we perform experiments to investigate performance on the MATH dataset. We find that our AMPS pretraining dataset increases MATH accuracy by approximately as much as a $15\times$ increase in model size, and that adding gigabytes of mathematics pretraining data from Math StackExchange does not help. We also find that models are highly overconfident. The models can also learn to reliably generate LaTeX step-by-step solutions and even graphical figures, even though the steps in the generated solutions are currently dubious. We observe that training on MATH step-by-step solutions also improves accuracy. Overall we find that MATH accuracy is increasing, but it is far from the ceiling.

3.1 Analyzing Model Performance

Model Size. While increasing model parameters often automatically solves many tasks (Brown et al., 2020), we find that MATH is unusually challenging for enormous Transformers. Table 1 shows that the average accuracy across subjects for the smallest model, GPT-2 with 0.1 billion parameters, is

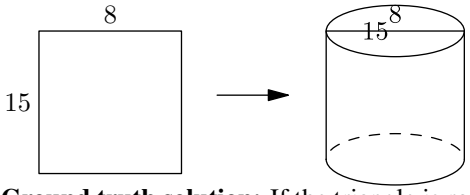
Problem: Suppose a and b are positive real numbers with $a > b$ and $ab = 8$. Find the minimum value of $\frac{a^2+b^2}{a-b}$.

Generated solution: Let $s = a + b$. By QM-AM, $\sqrt{\frac{a^2+b^2}{a-b}} \geq \frac{a+b}{2} = \frac{s}{2}$. Then $\frac{a^2+b^2}{a-b} \geq \frac{s^2}{4}$, so $a^2 + b^2 \geq \frac{s^2}{2}$. Hence, $\frac{a^2+b^2}{a-b} \geq \frac{s^2}{2} + \frac{2s^2}{a-b} = \frac{s^2}{2} + 2s^2 = \boxed{8}$.

Ground truth solution: We can write $\frac{a^2+b^2}{a-b} = \frac{a^2+b^2-2ab+16}{a-b} = \frac{(a-b)^2+16}{a-b} = a - b + \frac{16}{a-b}$. By AM-GM, $a - b + \frac{16}{a-b} \geq 2\sqrt{(a-b) \cdot \frac{16}{a-b}} = 8$. Equality occurs when $a - b = 4$ and $ab = 8$. We can solve these equations to find $a = 2\sqrt{3} + 2$ and $b = 2\sqrt{3} - 2$. Thus, the minimum value is $\boxed{8}$.

Problem: Right ΔABC has legs measuring 8 cm and 15 cm. The triangle is rotated about one of its legs. What is the number of cubic centimeters in the maximum possible volume of the resulting solid? Express your answer in terms of π.

Generated solution: Rotating the triangle about one leg gives us the following triangle:



$$\frac{(8)(15)}{(8)(-1)+(15)(-1)} = \frac{120}{7} = \boxed{120\pi}.$$

Ground truth solution: If the triangle is rotated about the shorter leg, then the radius is the longer leg and the height is the shorter leg, and the volume is $\frac{1}{3} \cdot (15^2\pi)(8) = 600\pi$ cubic centimeters. If the triangle is rotated about the longer leg, then the radius is the shorter leg and the height is the longer leg, and the volume is $\frac{1}{3}(8^2\pi)(15)$, which is $\frac{8}{15}$ of the volume we found earlier. So, the maximum possible volume is $\boxed{600\pi}$ cubic centimeters.

Figure 2: Problems, step-by-step solutions generated by our GPT-2 1.5B model, and ground truth solutions. Observe that models can provide the right answer yet generate a misleading and wrong explanation. The second generated solution demonstrates that models are capable of generating asymptote commands to create figures and graphics.

5.4%. Meanwhile, a GPT-2 model with $15\times$ the number of parameters, attains 6.9% accuracy, a 28% relative improvement. This indicates that while having more parameters helps, absolute accuracy remains far from the ceiling and is only increasing slowly, quite unlike most other text-based tasks.

AMPS Pretraining. As an ablation, we test how performance of models with AMPS pretraining compares with models that were not pretrained on AMPS. Without pretraining on AMPS, a GPT-2 (1.5B) model fine-tuned on MATH attains 5.5% accuracy. In contrast, a GPT-2 (1.5B) model both pretrained on AMPS and fine-tuned on MATH attains 6.9%, a 25% relative improvement in accuracy. Consequently AMPS increases accuracy about as much as a $15\times$ increase in parameters, indicating its value as a pretraining dataset.

We tried additionally pretraining on StackExchange, a real-world but less curated source of mathematics text. A GPT-2 (0.3B) model pretrained on both AMPS and questions and answers from Math StackExchange ($\sim$ 3 GB) had 6.0% accuracy, which is actually less than the 6.2% accuracy attained by pretraining on AMPS alone. Thus our dataset is more useful for pretraining even than diverse real-world mathematics data.

4 Conclusion

In this paper, We introduced the MATH benchmark, which enables the community to measure mathematical problem-solving ability. In addition to having answers, all MATH problems also include answer explanations, which models can learn from to generate their own step-by-step solutions. We also introduce AMPS, a diverse pretraining corpus that can enable future models to learn virtually all of K-12 mathematics. While most other text-based tasks are already nearly solved by enormous Transformers, MATH is fortunately different. We showed that accuracy is slowly increasing and, if trends continue, the community will need to discover conceptual and algorithmic breakthroughs to attain strong performance on MATH. Given the broad reach and applicability of mathematics, solving the MATH dataset with machine learning would be of profound practical and intellectual significance.

REFERENCES

T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, T. Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964, 2020.

T. Henighan, J. Kaplan, Mor Katz, Mark Chen, Christopher Hesse, J. Jackson, Heewoo Jun, T. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, B. Mann, A. Radford, Aditya Ramesh, Nick Ryder, D. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *ArXiv*, abs/2010.14701, 2020.

George Pólya. How to solve it. 1945.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Eugene Wigner. The unreasonable effectiveness of mathematics in the natural sciences. 1960.