

---

# 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

## A Appendix

In this appendix, we have more comparisons with previous datasets, a discussion of logic and intelligence tests, further AMPS and MATH details, an analysis of model performance as difficulty level changes, and results with the BART architecture.

### A.1 Expanded Dataset Comparisons

We compared to ten datasets in the main paper, and now we will further describe plug-and-chug datasets. Dolphin18K (Huang et al., 2016) is one of the first modern datasets in this space and is based on Yahoo! Answers and includes questions such as “help!!!!!!(please) i cant figure this out!? what is the sum of  $4\frac{2}{5}$  and  $17\frac{3}{7}$  ?”. MathQA (Amini et al., 2019) builds on AQuA-RAT (Ling et al., 2017) and claims AQuA-RATs “rationales are noisy, incomplete and sometimes incorrect.” MathQA then cleans AQuA-RAT, though cleaning led the dataset size to be reduced by half of an order of magnitude. Miao et al. (2020) analyze MathQA and observe “the annotated formulas of 27% of the problems do not match their labeled answers,” and they obtain 86% accuracy on a cleaned version of MATH-QA. In contrast AMPS is large and clean as questions are algorithmically generated, and our MATH dataset is carefully curated by the competition mathematics community and contains competition-level problems that are difficult.

### A.2 Logic and Intelligence Tests

While enormous Transformers perform poorly on MATH, they do well on other logic and intelligence tests.

We analyze Transformers on LogiQA (Liu et al., 2020), a task with logical reasoning questions such as “David knows Mr. Zhang’s friend Jack, and Jack knows David’s friend Ms. Lin. Everyone of them who knows Jack has a master’s degree, and everyone of them who knows Ms. Lin is from Shanghai. Who is from Shanghai and has a master’s degree?” As shown in Figure 1, Transformers are improving on LogiQA, so much so that they will attain human-level performance relatively soon, should trends continue.

We also find that Transformers also do well on the C-Test, a pattern completion test that has a 77% correlation with human IQ (Hernández-Orallo, 2000). An example of a problem from C-Test is the sequence “a, a, z, c, y, e, x, \_” which has the answer “g.” We regenerated hundreds of C-Test examples to test GPT-3 (175B) in a 5-shot setting. While GPT-3 had abysmal performance when the sequences were letters, converting letters to numbers helped. After changing ‘a’ to 0, ‘b’ to 1, . . . , and ‘z’ to 25, accuracy became approximately 40% on the hardest examples (C-Test questions with complexity “13”). For comparison, on these same examples, average humans attained around 20% accuracy (Hernández-Orallo, 2000).

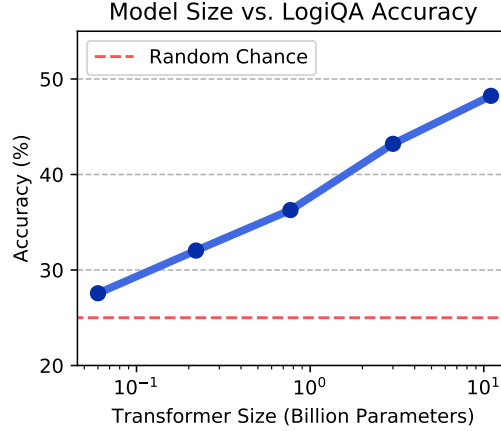


Figure 1: Difficult natural language tasks such as LogiQA will soon be solved just by making models larger, assuming trends continue. The Transformers in this figure are UnifiedQA (Khashabi et al., 2020) models of various sizes.

```
size(40);
draw(shift(1.38,0)*yscale(0.3)*Circle((0,0),.38));

draw((1,0)--(1,-2));
draw((1.76,0)--(1.76,-2));

draw((1,-2)..(1.38,-2.114)..(1.76,-2));
path p=(1.38,-2.114)..(1.74,-1.5)..(1,-0.5)..(1.38,-.114);
pair a=(1.38,-2.114), b=(1.76,-1.5);
path q=subpath(p,1,2);
path r=subpath(p,0,1);
path s=subpath(p,2,3);
draw(r);
draw(s);
draw(q,dashed);

label("$5$",midpoint((1.76,0)--(1.76,-2)),E);
```

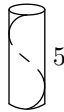


Figure 2: Example of asymptote code and the figure it produces.

### A.3 Further Dataset Information

**Rendering Graphics.** For the first time, our dataset makes it possible for text-based models to process graphical mathematical figures by expressing figures in asymptote code. For example, Figure 2 shows asymptote code and the figure it produces. In short, it is possible to concisely specify many visual mathematics problems with code, sidestepping the complexity of multi-modal models.

**AMPS Examples.** We show concrete examples from AMPS in Figure 3. AMPS is a mixture of examples from Khan Academy and our 100 Mathematica modules.

**Contrasting AMPS and DeepMind Mathematics.** AMPS has several hundred exercise types or modules (Khan Academy has 693 modules and Mathematica has 100), while DeepMind mathematics (DM) has only a few dozen. We show all Khan Academy modules in Figures 7 to 10. Most DM exercises increase the diversity of problems by simply having a wide range of coefficients and constants. For example, its derivatives module exclusively covers polynomial derivatives with wide-

Example from a Khan Academy [module](#):

Problem: In history class, the girl to boy ratio is 9 to 6. If there are a total of 60 students, how many boys are there?

Solution: A ratio of 9 girls to 6 boys means that a set of 15 students will have 9 girls and 6 boys. A class of 60 students has 4 sets of 15 students. Because we know that there are 6 boys in each set of 15 students, the class must have 4 groups of 6 boys each. There is a total of 24 boys in history class.

---

Example Mathematica code that generates practice problems:

```
In[1]:= For[i=0,i<50000,i++,
roundbasis = RandomChoice[{0.8,0.1,0.05,0.05}->{1,1/2,1/3,1/5}];
d1 = RandomInteger[{1,6}];
d2 = RandomInteger[{1,3}];
q=0;
p=0;
While[q==0,
For[j=0,j<d1,j++,
q += Round[RandomReal[{-5,5}], roundbasis]*x^j;
];
];
While[p==0,
For[j=0,j<d2,j++,
p += Round[RandomReal[{-5,5}], roundbasis]*x^j;
];
];
p = RandomChoice[{p,Expand[q*p]}];
Export["/amps/mathematica/algebra/polynomial_gcd/"<>ToString[i]<>".txt",
{"Problem:\nFind the greatest common divisor of $"
<>ToString[TeXForm[p//TraditionalForm]]<> "$ and $"
<>ToString[TeXForm[q//TraditionalForm]]<> "$.",
"Answer:\n$" <>ToString[TeXForm[PolynomialGCD[p,q]//TraditionalForm]]<> "$"}]
]
```

Figure 3: A Khan Academy problem and solution, followed by the code for a simple Mathematica module used to generate polynomials GCD problems. These problems are available in AMPS.

ranging coefficients, while ours covers mixtures of dozens of major analytic functions. DM opts not to cover concepts and subjects such as logarithms and geometry, unlike AMPS. While DM is formatted in plaintext, AMPS is formatted in  $\text{\LaTeX}$ . Finally, while DM solely has final answers, all 693 Khan Academy modules and 37 of our Mathematica modules have full step-by-step solutions.

#### A.4 Difficulty Analysis

We break down MATH accuracy by difficulty levels. In Figure 4, we observe that human difficulty and machine difficulty track each other. In Figure 5, we find that accuracy can vary by level and subject substantially. Finally, in Figure 6a and Figure 6b, we analyze the relation between accuracy and problem and solution length, and find that problems with long questions or ground truth solutions indeed tend to be more difficult than problems with short questions or solutions.

#### A.5 Results with the BART Architecture

We use BART (Lewis et al., 2020) to determine whether other existing architectures can improve performance. In the main paper we analyzed the performance of various GPT models, which are unidirectional decoder models. Lewis et al. (2020) introduce BART, which has a bidirectional encoder and unidirectional decoder. While T5 has a similar architecture to BART, its tokenizer removes  $\text{\LaTeX}$  symbols, while BART’s tokenizer does not. Hence we use BART in this paper. After pretraining BART-Large (0.4B) on AMPS and fine-tuning BART on MATH, we find that it obtains 4.9% on

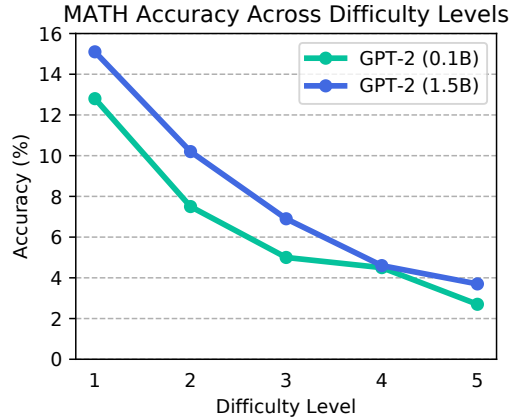


Figure 4: Problems that are more difficult for humans are also more difficult for GPT-2.

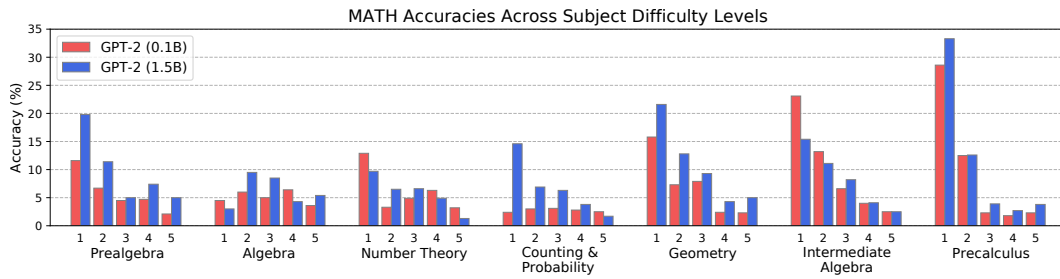


Figure 5: Accuracy per subject per difficulty level.

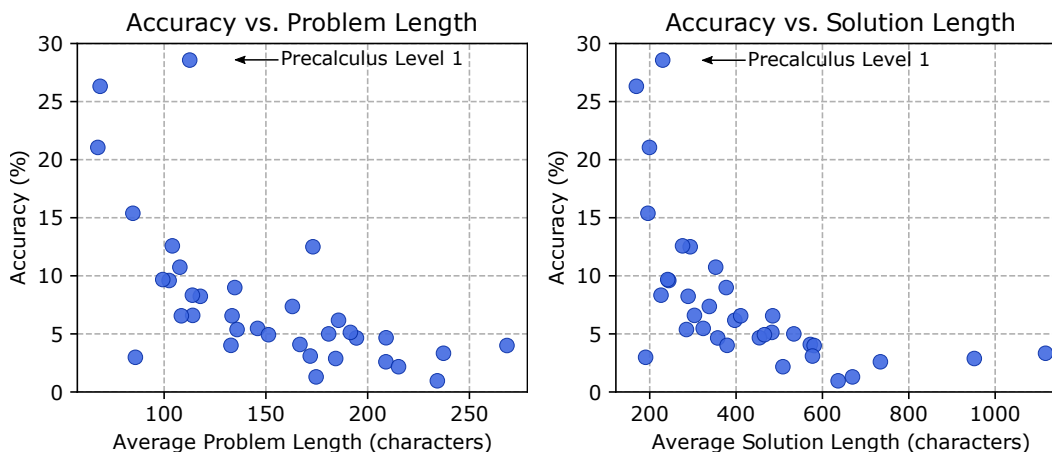
MATH’s test set, which is slightly worse than our smallest GPT-2 model. Consequently models with a bidirectional encoder and unidirectional decoder do not yield marked changes in MATH accuracy.

## A.6 Further Human Evaluation Details

Because MATH requires a strong mathematical background to perform well on, and a long amount of time to solve problems, we were restricted to assessing six human participants and could not rely on crowdsourcing sites such as Amazon Mechanical Turk. All participants are university students studying computer science. Four of the participants are authors on the paper. The other two participants are friends or acquaintances of the authors, as this survey was only to give a rough sense of human-level performance. All participants gave consent to use their name, though we opt not to include it. Participants had one hour to complete the questions, all without using a calculator. Participants were offered remuneration, though they all offered to volunteer to work on the problems. The instructions and questions used are as follows.

“This research study is being conducted by the Steinhardt Group at UC Berkeley. For questions about this study, please contact Dan Hendrycks at [hendrycks@berkeley.edu](mailto:hendrycks@berkeley.edu). In this study, you will have sixty minutes to complete twenty mathematics problems, all without using a calculator. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.”

1. A 6-sided die is weighted so that the probability of any number being rolled is proportional to the value of the roll. (So, for example, the probability of a 2 being rolled is twice that of a 1 being rolled.) What is the expected value of a roll of this weighted die? Express your answer as a common fraction.
2. The square of 15 is 225. The square of what other number is 225?
3. Find the sum of all values of  $x$  such that  $|x - 1| = 7$ .
4. The parabolas defined by the equations  $y = -x^2 - x + 1$  and  $y = 2x^2 - 1$  intersect at points  $(a, b)$  and  $(c, d)$ , where  $c \geq a$ . What is  $c - a$ ? Express your answer as a common fraction.



(a) Subject accuracy vs problem length. Each point represents a subject at a specific difficulty level. We exclude problems with asymptote figures. Results are from GPT-2 (1.5B).

(b) Subject accuracy vs solution length. Each point represents a subject at a specific difficulty level. We exclude problems with asymptote figures. Results are from GPT-2 (1.5B).

5. If  $a = 8$ , what is the value of  $\left(16\sqrt[3]{a^2}\right)^{\frac{1}{3}}$ ?
6. Let  $p(x)$  be a cubic polynomial such that  $p(2) = 0$ ,  $p(-1) = 0$ ,  $p(4) = 6$ , and  $p(5) = 8$ . Find  $p(7)$ .
7. Let  $S$  be the set of complex numbers of the form  $a + bi$ , where  $a$  and  $b$  are integers. We say that  $z \in S$  is a unit if there exists a  $w \in S$  such that  $zw = 1$ . Find the number of units in  $S$ .
8. Find the remainder when  $1 + 2 + 2^2 + 2^3 + \cdots + 2^{100}$  is divided by 7.
9. The length of a rectangle is  $3x + 10$  feet and its width is  $x + 12$  feet. If the perimeter of the rectangle is 76 feet, how many square feet are in the area of the rectangle?
10. A European train compartment has six seats. Four of the seats are broken. Wilhelm needs to fill out a form to indicate that there are broken seats. If he randomly checks off four of the seats in the diagram, what is the probability that he marked the correct seats? Express your answer as a common fraction.
11. We have a triangle  $\triangle ABC$  where  $AC = 17$ ,  $BC = 15$ , and  $AB = 8$ . Let  $M$  be the midpoint of  $AB$ . What is the length of  $CM$ ?
12. If  $n$  gives a remainder of 3 when divided by 7, then what remainder does  $2n + 1$  give when divided by 7?
13. Our club has 25 members, and wishes to pick a president, secretary, and treasurer. In how many ways can we choose the officers, if individual members are allowed to hold 2, but not all 3, offices?
14. Find the minimum possible value of

$$\sqrt{58 - 42x} + \sqrt{149 - 140\sqrt{1 - x^2}}$$

where  $-1 \leq x \leq 1$ ?

15. Let  $a$ ,  $b$ , and  $c$  be the roots of  $x^3 + 7x^2 - 11x - 2 = 0$ . Find  $a + b + c$ .
16. Let  $\mathcal{H}$  be the hyperbola with foci at  $(\pm 5, 0)$  and vertices at  $(\pm 3, 0)$ , and let  $\mathcal{C}$  be the circle with center  $(0, 0)$  and radius 4. Given that  $\mathcal{H}$  and  $\mathcal{C}$  intersect at four points, what is the area of the quadrilateral formed by the four points?
17. If  $f(x) = x^2 - 2x + 1$  and  $g(x) = \sqrt{2x + 1}$  what is the value of  $f(g(4)) - g(f(3))$ ?
18. Find the value of  $r$  such that  $\frac{6r^2 - 19r - 7}{2r - 7} = 4r - 3$ .
19. For  $x > 0$ , the area of the triangle with vertices  $(0, 0)$ ,  $(x, 0)$  and  $(x, 5)$  is 30 square units. What is the value of  $x$ ?
20. Find the units digit of the following within the indicated number base:  $413_6 - 215_6$ .

**Khan Academy Modules (1/4):** 2 step equations; 2-step addition word problems within 100; 2-step subtraction word problems within 100; 2-step word problems; absolute minima and maxima (closed intervals); absolute minima and maxima (entire domain); absolute value equations; absolute value of complex numbers; add and subtract complex numbers; add and subtract matrices; add and subtract polynomials; add and subtract rational expressions; add and subtract rational expressions: factored denominators; add and subtract rational expressions: like denominators; add and subtract rational expressions: unlike denominators; add and subtract vectors; add 1 or 10; add 1s or 10s (no regrouping); add 3 numbers; add and subtract fractions; add and subtract fractions word problems; add and subtract within 20 word problems; add fractions with unlike denominators; add within 10; add within 1000; add within 20; add within 5; adding and subtracting decimals word problems; adding and subtracting in scientific notation; adding and subtracting negative fractions; adding and subtracting negative numbers; adding and subtracting rational numbers; adding and subtracting decimals word problems; adding and subtracting fractions; adding and subtracting mixed numbers 0.5; adding and subtracting mixed numbers 1; adding and subtracting polynomials; adding and subtracting radicals; adding and subtracting rational expressions 0.5; adding and subtracting rational expressions 1; adding and subtracting rational expressions 1.5; adding and subtracting rational expressions 2; adding and subtracting rational expressions 3; adding and subtracting rational numbers; adding and subtracting with unlike denominators 5; adding and subtracting with unlike denominators 6; adding decimals (hundredths); adding decimals (tenths); adding decimals and whole numbers (hundredths); adding decimals and whole numbers (tenths); adding decimals: thousandths; adding fractions; adding fractions 0.5; adding up to four 2-digit numbers; adding vectors; addition and subtraction word problems; addition and subtraction word problems 2; addition word problems within 100; age word problems; amplitude of sinusoidal functions from equation; analyze concavity; angle addition postulate; angle of complex numbers; approximation with local linearity; arc length; area and perimeter of rectangles word problems; area between two curves; area between two curves given end points; area between two polar curves; area bounded by polar curves; area bounded by polar curves intro; area of a circle; area of parallelograms; area problems; areas of circles and sectors; arithmetic sequences 1; arithmetic sequences 2; arithmetic series; average value of a function; average word problems; basic division; basic multiplication; basic partial derivatives; basic set notation; binomial probability formula; calculating binomial probability; center and radii of ellipses from equation; chain rule capstone; chain rule intro; change of variables: bound; change of variables: factor; circles and arcs; circulation form of green's theorem; classifying critical points; combinations; combined vector operations; combining like terms; combining like terms with distribution; combining like terms with negative coefficients; combining like terms with rational coefficients; complementary and supplementary angles; complete solutions to 2-variable equations; completing the square; completing the square (intermediate); completing the square (intro); complex numbers from absolute value and angle; complex plane operations; composite exponential function differentiation; composite numbers; conditional statements and truth value; construct exponential models; construct sinusoidal functions; continuity at a point (algebraic); converting between point slope and slope intercept form; converting between slope intercept and standard form; converting decimals to fractions 1; converting decimals to fractions 2; converting decimals to percents; converting fractions to decimals; converting mixed numbers and improper fractions; converting multi digit repeating decimals to fractions; converting multi-digit repeating decimals to fractions; converting percents to decimals; converting recursive and explicit forms of arithmetic sequences; converting recursive and explicit forms of geometric sequences; counting 1; counting 2; cube roots; cube roots 2; cumulative geometric probability; defined and undefined matrix operations; definite integral as the limit of a riemann sum; definite integrals of piecewise functions; definite integrals: common functions; definite integrals: reverse power rule; degrees to radians; density word problems; dependent probability; derivatives 1; derivatives of  $a^x$  and  $\log_a x$ ; derivatives of  $\sin(x)$  and  $\cos(x)$ ; derivatives of  $\tan(x)$ ,  $\cot(x)$ ,  $\sec(x)$ , and  $\csc(x)$ ; derivatives of  $e^x$  and  $\ln(x)$ ; determinant of a 2x2 matrix; determinant of a 3x3 matrix; difference of squares; differentiability at a point: algebraic; differential equations: exponential model equations; differentiate integer powers (mixed positive and negative); differentiate polynomials; differentiate products; differentiate quotients; differentiate rational functions; differentiate related functions; differentiating using multiple rules; direct comparison test; direct substitution with limits that don't exist; direction of vectors; disc method: revolving around other axes; disc method: revolving around x- or y-axis; discount, markup, and commission word problems; discount, tax, and tip word problems; disguised derivatives; distance between point and line; distance formula; distributive property with variables; divide by 1; divide by 10; divide by 2; divide by 3; divide by 4; divide by 5; divide by 6; divide by 7; divide by 8; divide by 9; divide complex numbers; divide decimals by whole numbers; ...

Figure 7: Khan Academy modules in our AMPS pretraining dataset (Part 1).

**Khan Academy Modules (2/4):** divide fractions by whole numbers; divide mixed numbers; divide polynomials by linear expressions; divide polynomials by monomials (with remainders); divide polynomials by  $x$  (no remainders); divide polynomials by  $x$  (with remainders); divide polynomials with remainders; divide powers; divide quadratics by linear expressions (no remainders); divide quadratics by linear expressions (with remainders); divide whole numbers by 0.1 or 0.01; divide whole numbers by decimals; divide whole numbers by fractions; divide whole numbers to get a decimal (1-digit divisors); divide whole numbers to get a decimal (2-digit divisors); divide with remainders (2-digit by 1-digit); dividing complex numbers; dividing decimals 1; dividing decimals 2; dividing decimals: hundredths; dividing decimals: thousandths; dividing fractions; dividing fractions word problems; dividing fractions word problems 2; dividing mixed numbers with negatives; dividing negative numbers; dividing polynomials by binomials 1; dividing polynomials by binomials 2; dividing polynomials by binomials 3; dividing positive and negative fractions; dividing positive fractions; dividing rational numbers; dividing whole numbers by fractions; dividing whole numbers by unit fractions; dividing whole numbers like  $56/35$  to get a decimal; divisibility 0.5; divisibility tests; domain of a function; double integrals with variable bounds; empirical rule; equation of a circle in factored form; equation of a circle in non factored form; equation of a hyperbola; equation of a parabola from focus and directrix; equation of an ellipse; equation of an ellipse from features; equations and inequalities word problems; equations of parallel and perpendicular lines; equations with parentheses; equations with parentheses: decimals and fractions; equations with variables on both sides; equations with variables on both sides: decimals and fractions; equivalent fractions; estimating square roots; evaluate composite functions; evaluate function expressions; evaluate functions; evaluate logarithms; evaluate logarithms (advanced); evaluate logarithms: change of base rule; evaluate piecewise functions; evaluate radical expressions challenge; evaluate sequences in recursive form; evaluating composite functions; evaluating expressions in 2 variables; evaluating expressions in one variable; evaluating expressions with multiple variables; evaluating expressions with multiple variables: fractions and decimals; evaluating expressions with one variable; evaluating expressions with variables word problems; evaluating logarithms; evaluating logarithms 2; expected value; explicit formulas for arithmetic sequences; explicit formulas for geometric sequences; exponent rules; exponential expressions word problems (algebraic); exponential model word problems; exponential vs. linear growth over time; exponents with integer bases; exponents with negative fractional bases; expressing ratios as fractions; expressions with unknown variables; expressions with unknown variables 2; extend arithmetic sequences; extend geometric sequences; extend geometric sequences: negatives and fractions; extraneous solutions to rational equations; factor higher degree polynomials; factor polynomials using structure; factor quadratics by grouping; factor using polynomial division; factor with distributive property (variables); factoring difference of squares 1; factoring difference of squares 2; factoring difference of squares 3; factoring linear binomials; factoring polynomials by grouping; factoring polynomials with two variables; factoring quadratics 1; factoring quadratics with a common factor; features of a circle from its expanded equation; features of a circle from its standard equation; features of quadratic functions; find area elements; find composite functions; find critical points; find critical points of multivariable functions; find inflection points; find inverses of rational functions; find missing divisors and dividends (1-digit division); find missing factors (1-digit multiplication); find missing number (add and subtract within 20); find the inverse of a  $2 \times 2$  matrix; find the missing number (add and subtract within 1000); find trig values using angle addition identities; finding absolute values; finding curl in 2d; finding curl in 3d; finding derivative with fundamental theorem of calculus; finding derivative with fundamental theorem of calculus: chain rule; finding directional derivatives; finding divergence; finding gradients; finding inverses of linear functions; finding partial derivatives; finding percents; finding perimeter; finding tangent planes; finding the laplacian; finite geometric series; finite geometric series word problems; foci of an ellipse from equation; fraction word problems 1; fractional exponents; fractional exponents 2; fractions as division by a multiple of 10; function as a geometric series; function inputs and outputs: equation; function rules from equations; gcf and lcm word problems; general triangle word problems; geometric probability; geometric sequences 1; geometric sequences 2; geometric series formula; graphing points and naming quadrants; graphing systems of equations; greatest common factor; greatest common factor of monomials; higher order partial derivatives; identify composite functions; identify separable equations; identifying numerators and denominators; identifying slope of a line; imaginary unit powers; implicit differentiation; improper integrals; increasing and decreasing intervals; indefinite integrals:  $e^x$  and  $1/x$ ; indefinite integrals:  $\sin$  and  $\cos$ ; independent probability; inequalities word problems; infinite geometric series; integer sums; integral test; integrals and derivatives of functions with known power series; integrals in spherical and cylindrical coordinates; ...

Figure 8: Khan Academy modules in AMPS (Part 2).

**Khan Academy Modules (3/4):** integrate and differentiate power series; integrating trig functions; integration by parts; integration by parts: definite integrals; integration using completing the square; integration using long division; integration using trigonometric identities; integration with partial fractions; intercepts from an equation; interpret quadratic models; interval of convergence; inverse of a 3x3 matrix; inverses of functions; iterated integrals; jacobian determinant; l'hôpital's rule (composite exponential functions); l'hôpital's rule:  $0/0$ ; l'hôpital's rule:  $\infty/\infty$ ; lagrange error bound; least common multiple; limits at infinity of quotients; limits at infinity of quotients with square roots; limits at infinity of quotients with trig; limits by direct substitution; limits by factoring; limits of piecewise functions; limits of trigonometric functions; limits using conjugates; limits using trig identities; line integrals in vector fields; linear equation and inequality word problems; linear equations with unknown coefficients; linear equations word problems; linear models word problems; logical arguments and deductive reasoning; maclaurin series of  $\sin(x)$ ,  $\cos(x)$ , and  $e^x$ ; make 10; manipulate formulas; markup and commission word problems; matrix addition and subtraction; matrix dimensions; matrix elements; matrix equations: addition and subtraction; matrix equations: scalar multiplication; matrix row operations; matrix transpose; mean, median, and mode; midline of sinusoidal functions from equation; midpoint of a segment; miscellaneous; model with one-step equations and solve; modeling with multiple variables; modeling with sinusoidal functions; modeling with sinusoidal functions: phase shift; motion along a curve (differential calc); motion problems (differential calc); motion problems (with integrals); multi-digit addition; multi-digit division; multi-digit multiplication; multi-digit subtraction; multi-step linear inequalities; multi-step word problems with whole numbers; multiplication and division word problems; multiplication and division word problems (within 100); multiply and divide complex numbers in polar form; multiply and divide powers (integer exponents); multiply and divide rational expressions (advanced); multiply binomials; multiply binomials by polynomials; multiply binomials intro; multiply by 0 or 1; multiply by 2 and 4; multiply by 5 and 10; multiply by tens word problems; multiply complex numbers; multiply decimals (1 and 2-digit factors); multiply decimals (up to 4-digit factors); multiply difference of squares; multiply matrices; multiply matrices by scalars; multiply mixed numbers; multiply monomials; multiply monomials by polynomials; multiply powers; multiply unit fractions and whole numbers; multiply whole numbers and decimals; multiplying and dividing in scientific notation; multiplying a matrix by a matrix; multiplying a matrix by a vector; multiplying and dividing complex numbers in polar form; multiplying and dividing negative numbers; multiplying and dividing rational expressions 1; multiplying and dividing rational expressions 2; multiplying and dividing rational expressions 3; multiplying and dividing rational expressions 4; multiplying and dividing rational expressions 5; multiplying and dividing scientific notation; multiplying by multiples of 10; multiplying complex numbers; multiplying decimals like  $0.847 \times 3.54$  (standard algorithm); multiplying decimals like  $2.45 \times 3.6$  (standard algorithm); multiplying decimals like  $4 \times 0.6$  (standard algorithm); multiplying expressions 1; multiplying fractions; multiplying fractions by integers; multiplying mixed numbers 1; multiplying negative numbers; multiplying polynomials; multiplying polynomials 0.5; multiplying positive and negative fractions; multiplying rational numbers; multivariable chain rule; multivariable chain rule intro; negative exponents; new operator definitions 1; new operator definitions 2; normal form of green's theorem; number of solutions of quadratic equations; one step equations; one step equations with multiplication; one-step addition and subtraction equations; one-step addition and subtraction equations: fractions and decimals; one-step equations with negatives (add and subtract); one-step equations with negatives (multiply and divide); one-step inequalities; one-step multiplication and division equations; one-step multiplication and division equations: fractions and decimals; operations with logarithms; order of operations; order of operations (no exponents); order of operations 2; order of operations challenge; order of operations with negative numbers; ordered pair solutions to linear equations; p-series; parametric curve arc length; parametric equations differentiation; parametric velocity and speed; partial derivatives of vector valued functions; partial fraction expansion; partial sums intro; particular solutions to differential equations; particular solutions to separable differential equations; parts of complex numbers; percent problems; perfect squares; period of sinusoidal functions from equation; permutations; permutations and combinations; planar motion (differential calc); planar motion (with integrals); polar and rectangular forms of complex numbers; polynomial special products: difference of squares; polynomial special products: perfect square; positive and zero exponents; positive exponents with positive and negative bases; potential functions; power rule (negative and fractional powers); power rule (positive integer powers); power rule (with rewriting the expression); powers of complex numbers; powers of fractions; powers of powers; prime numbers; probabilities of compound events; probability 1; probability in normal density curves; probability of "at least one" success; probability with permutations and combinations; problems involving definite integrals (algebraic); ...

Figure 9: Khan Academy modules in AMPS (Part 3).



**Khan Academy Modules (4/4):** properties of exponents (rational exponents); proportion word problems; pythagorean identities; pythagorean theorem; quadratic word problems (factored form); quadratic word problems (standard form); quadratic word problems (vertex form); quadratics by factoring; quadratics by taking square roots; radians and degrees; radians to degrees; radical equations; radius, diameter, and circumference; range of a function; rate conversion; rate problems; rate problems 2; rates of change in other applied contexts (non-motion problems); rates with fractions; ratio test; ratio word problems; reciprocal trig functions; recursive formulas for arithmetic sequences; recursive formulas for geometric sequences; regroup when adding 1-digit numbers; relate addition and subtraction; related rates (advanced); related rates (multiple rates); related rates (pythagorean theorem); related rates intro; relationship between exponentials and logarithms; relative minima and maxima; remainder theorem; remainder theorem and factors; removable discontinuities; represent linear systems with matrices; represent linear systems with matrix equations; reverse power rule; reverse power rule: negative and fractional powers; reverse power rule: rewriting before integrating; reverse power rule: sums and multiples; rewriting decimals as fractions challenge; right triangle trigonometry word problems; roots of decimals and fractions; sample and population standard deviation; scalar matrix multiplication; scalar multiplication; scientific notation; secant lines and average rate of change; secant lines and average rate of change with arbitrary points; secant lines and average rate of change with arbitrary points (with simplification); second derivative test; second derivatives (implicit equations); second derivatives (parametric functions); second derivatives (vector-valued functions); segment addition; separable differential equations; significant figures; simplify roots of negative numbers; simplify square roots (variables); simplify square-root expressions; simplifying expressions with exponents; simplifying fractions; simplifying radicals; simplifying radicals 2; simplifying rational expression with exponent properties; simplifying rational expressions 2; simplifying rational expressions 3; simplifying rational expressions 4; sinusoidal models word problems; slope-intercept from two points; solid geometry; solutions to quadratic equations; solutions to systems of equations; solve equations using structure; solve exponential equations using exponent properties; solve exponential equations using exponent properties (advanced); solve exponential equations using logarithms: base-10 and base-e; solving equations in terms of a variable; solving for the x intercept; solving for the y intercept; solving proportions; solving quadratics by completing the square 1; solving quadratics by completing the square 2; solving quadratics by factoring; solving quadratics by factoring 2; solving quadratics by taking the square root; solving rational equations 1; solving rational equations 2; special right triangles; square and cube challenge; square roots of perfect squares; standard deviation; standard deviation of a population; stokes' theorem; substitution with negative numbers; subtract decimals (hundredths); subtract decimals and whole numbers (hundredths); subtract within 10; subtract within 1000; subtract within 20; subtract within 5; subtracting decimals (tenths); subtracting decimals and whole numbers (tenths); subtracting decimals: thousandths; subtracting fractions; subtracting fractions with common denominators; subtracting fractions with unlike denominators; subtraction word problems within 100; summation notation intro; sums of consecutive integers; surface integrals to find surface area; switching bounds on double integrals; symbols practice: the gradient; systems of equations; systems of equations with elimination; systems of equations with simple elimination; systems of equations with substitution; systems of equations word problems; tangents to polar curves; taylor and maclaurin polynomials; the derivative and tangent line equations; the divergence theorem; the fundamental theorem of calculus and definite integrals; the hessian matrix; translate one-step equations and solve; trigonometry 0.5; trigonometry 1; trigonometry 1.5; trigonometry 2; triple integrals; two-step equations; two-step equations with decimals and fractions; two-step equations word problems; u-substitution: definite integrals; u-substitution: indefinite integrals; unit circle; unit vectors; use arithmetic sequence formulas; use geometric sequence formulas; use the properties of logarithms; use the pythagorean identity; using the mean value theorem; using the quadratic formula; using units to solve problems; variance; vector word problems; vector-valued functions differentiation; verify solutions to differential equations; vertex of a parabola; volume word problems; volumes with cross sections: squares and rectangles; volumes with cross sections: triangles and semicircles; washer method: revolving around other axes; washer method: revolving around x- or y-axis; word problems with "more" and "fewer" 2; write common decimals as fractions; write common fractions as decimals; write decimals as fractions; write differential equations; write equations of parallel and perpendicular lines; writing basic expressions with variables; writing basic expressions word problems; writing expressions; writing expressions 2; writing expressions with variables; writing expressions word problems; writing functions with exponential decay; writing linear functions word problems; writing proportional equations; writing proportions; wrong statements in triangle proofs; z scores 1; z scores 2; z scores 3; zero product property.

Figure 10: Khan Academy modules in AMPS (Part 4).

## B Checklist Information

**Legal Compliance.** We create and collect various mathematics problems to create MATH and AMPS.

AMPS consists of problems generated with Mathematica and Khan Academy code. Mathematica serves as a calculator and does not copyright its numerical answer outputs, in much the same way that other calculators do not copyright computations such as  $5^2 \pmod{2}$ . Khan Academy’s exercise framework follows an MIT License. Since we provide attribution, reuse is not restrictive save for attribution requirements.

MATH problems are created by the Mathematical Association of America (MAA). Although we do not commercialize MATH, we should like to demonstrate that we are far from the boundary for action or infringement. For decades, the MAA has not protected its problem IP even from separate organizations which sell MAA problems, such as AoPS. Courts have ruled that this implies the IP rights are permanently forfeited. We raise this point only to demonstrate the extent to which our reuse for research is within the law, because even commercial reuse of MAA problems is within the law and commonplace. Even so, the MATH dataset is not sold and is likely to have no effect on the value of the original problems. This analysis would be pertinent in the hypothetical situation where Fair Use doctrine did not exist, but MATH and AMPS are covered by Fair Use.

For MATH and AMPS, we abide by Fair Use §107: “the fair use of a copyrighted work, including such use by ... scholarship, or research, is not an infringement of copyright”, where fair use is determined by “the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes” and “the effect of the use upon the potential market for or value of the copyrighted work.”

**Dataset Intended Uses.** We document the dataset within the paper and note that the dataset and code for reproducing results is available at <https://github.com/hendrycks/apps>. We do not intend for this dataset to train models that help students cheat on mathematics exams. We intend for others to use this dataset in order to better forecast reasoning capabilities.

**Author Statement and License.** We bear all responsibility in case of violation of rights. The MATH data, AMPS data, and our open source code are under an MIT license.

## References

- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. *ArXiv*, abs/1905.13319, 2019.
- J. Hernández-Orallo. Beyond the turing test. *Journal of Logic, Language and Information*, 9:447–466, 2000.
- D. Huang, Shuming Shi, Chin-Yew Lin, J. Yin, and W. Ma. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *ACL*, 2016.
- Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system, 2020.
- M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461, 2020.
- Wang Ling, Dani Yogatama, Chris Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*, 2017.
- J. Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*, 2020.
- Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In *ACL*, 2020.