Appendix

A Training Dataset Details

The two main data sources for our training dataset are arXiv papers and web pages that contain mathematics. Here we present additional details on how the data from each source was collected and processed.

A.1 arXiv

The arXiv dataset contains 2M arXiv papers up to February 2021, in LaTeX format. If multiple LaTeX files were present, they were concatenated. Comments were removed, and anything before the first section header or after an appendix/bibliography header was removed. The title and abstract of each paper were added to the document from the arXiv metadata. In order to retain high quality documents and maximize the information per token, papers were filtered out if they were longer than 75k tokens, had on average more than 0.6 tokens per character, had no \section headers, or ended up being empty after processing. The final arXiv dataset after processing includes 1.2M papers totalling 58GB of data.

A.2 Mathematical web pages

We started with a collection of web pages that included the string "<math" or "MathJax-Element-" in the raw HTML, which we used as our filter for pages that that include mathematical content. We considered pages as of January 2022. We then used several heuristics to process the pages. We found empirically that these are sufficient to extract most of the available mathematical content in either LaTeX format or ASCII-math format. The majority of the documents (about 80% of documents) have one of these two formats:

1. A majority of these HTML documents contain math in TeX or AsciiMath format inside tags of the form <script type="math/latex"> or <script type="math/asciimath">.
2. Another common appearance of LaTeX happens with <annotation encoding="application/x-tex"> tags inside <math> MathML blocks. We extract the content of these <annotation> blocks but do not include other content from inside the <math> blocks.

The remaining documents (about 20%) generally have math in MathML format, which we discarded. After extracting the content in any of the previous two forms, we removed all other content that was inside <math> or <span id=MathJax-Element-*> blocks, because these blocks often encode the MathML version of TeX or AsciiMath content. After filtering, processing, and selecting only English documents, the final dataset size is 60GB.

B Model and Training Procedure Details

We start with pretrained BaseModel models, and perform unsupervised finetuning on our technical dataset to obtain Minerva .

The models have context length 2048. They are trained with batch size 128 (except for the 540B model which was trained with batch size 32) and without dropout.

The learning rate schedule was reciprocal square-root decay, which continued the schedule of the pretrained models. The 8B model was pretrained for 1M steps and further trained for 600k additional unsupervised finetuning steps. The 62B model was pretrained for 520k steps and further trained for 400k additional unsupervised finetuning steps. The 540B model was pretrained for 257k steps and was further trained for 383k additional steps during unsupervised finetuning.

Finally, the learning rate was dropped 10x and all models were then trained for 4% additional steps. We note that these models had a significantly larger batch size during pretraining.

We used the `t5x` framework [49] and trained our models with v4 TPU on Google Cloud. The 8B model was trained for 14 days on a v4-128, the 62B model was trained for 17 days on a v4-512, and the 540B model was trained for 29 days on a v4-1024.

C MATH Evaluation Details

C.1 MATH Answer Normalization

Extracting and evaluating the correctness of answers to math questions is non-trivial because answers can often be presented in many different ways, both in terms of formatting (e.g. answers can be underlined, or surrounded by a box) and in terms of mathematical content (a large number can be equivalently represented as 1,000 or 1000, answers about currency potentially have the currency symbol attached to them, etc.). Here we describe how final answers are extracted and normalized. After normalization, answers are compared using SymPy (see below). Failing to normalize answers properly will typically lead to falsely identifying correct answers as incorrect ("false negatives"), and therefore to underestimate the model's accuracy.

We first extract the final answer from the full model response, which potentially includes chain-of-thought reasoning. In the few-shot prompt, we used the format `"Final Answer: The final answer is ANSWER. I hope it is correct."` for every final answer. We look for this pattern in the model output and extract `ANSWER`.

We then apply a normalization function to this answer, shown in Listing 1. In order to develop it we manually inspected ground truth targets, samples from `Minerva`, and samples from OpenAI davinci-002. We were especially careful to avoid changes in the format of the ground truth target that might produce false positives.

```python
SUBSTITUTIONS = [
    ('an ', ''), ('a ', ''), ('.$', '$'), ('\\$', ''), (r'\ ', ''),
    (' ', ''), ('mbox', 'text'), (',\\text{and}', ','),
    ('\\text{and}', ','), ('\\text{m}', '\\text{}')
]
REMOVED_EXPRESSIONS = [
    'square', 'ways', 'integers', 'dollars', 'mph', 'inches', 'ft',
    'hours', 'km', 'units', '\\ldots', 'sue', 'points', 'feet',
    'minutes', 'digits', 'cents', 'degrees', 'cm', 'gm', 'pounds',
    'meters', 'meals', 'edges', 'students', 'childrentickets', 'multiples',
    '\\text{s}', '\\text{.}', '\\text{\ns}', '\\text{}^2',
    '\\text{}^3', '\\text{\n}', '\\text{}', r'\mathrm{th}',
    r'^\circ', r'^{\circ}', r'\;', r',\!', '{,}', '"', '\\dots'
]

def normalize_final_answer(final_answer: str) -> str:
  """Normalize a final answer to a quantitative reasoning question."""
  final_answer = final_answer.split('=')[-1]

  for before, after in SUBSTITUTIONS:
    final_answer = final_answer.replace(before, after)
  for expr in REMOVED_EXPRESSIONS:
    final_answer = final_answer.replace(expr, '')

  # Extract answer that is in LaTeX math, is bold,
  # is surrounded by a box, etc.
  final_answer = re.sub(r'(.*?)(\$)(.*?)(\$)(.*)', '$\\3$', final_answer)
  final_answer = re.sub(r'(\\text\{)(.*?)(\})', '\\2', final_answer)
  final_answer = re.sub(r'(\\textbf\{)(.*?)(\})', '\\2', final_answer)
  final_answer = re.sub(r'(\\overline\{)(.*?)(\})', '\\2', final_answer)
  final_answer = re.sub(r'(\\boxed\{)(.*)(\})', '\\2', final_answer)

  # Normalize shorthand TeX:
  #  \fracab -> \frac{a}{b}
  #  \frac{abc}{bef} -> \frac{abc}{bef}
  #  \fracabc -> \frac{a}{b}c
  #  \sqrta -> \sqrt{a}
  #  \sqrtab -> sqrt{a}b
  final_answer = re.sub(
    r'(frac)([^{])(.)', 'frac{\\2}{\\3}', final_answer)
  final_answer = re.sub(
    r'(sqrt)([^{])', 'sqrt{\\2}', final_answer)
  final_answer = final_answer.replace('$', '')

  # Normalize 100,000 -> 100000
  if final_answer.replace(',', '').isdigit():
    final_answer = final_answer.replace(',', '')

  return final_answer
```

Listing 1: Python code used to normalize final answers.

After applying this normalization function, we checked whether the formatted target and prediction strings are `SymPy`-equivalent. `SymPy` equivalence is determined by parsing the answers via `sympy.parsing.latex.parse_latex` and then checking whether substracting the two resulting `SymPy` objects and applying `sympy.simplify` gives zero. We set a timeout of 5s when calling `sympy.simplify`, and labeled strings as nonequivalent if this timeout was exceeded.

For MATH problems, `SymPy` equivalence improved overall accuracy by around 1%. See Table 5 for the accuracies in MATH with only exact string match vs. `SymPy` equivalence.

Table 5: Comparing MATH accuracy when evaluating results with and without `SymPy` processing.

	MATH Accuracy	
	without `SymPy`	with `SymPy`
Minerva 8B	13.3	14.1
Minerva 8B Majority	24.6	25.4
Minerva 62B	26.5	27.6
Minerva 62B Majority	42.2	43.4
OpenAI davinci-002	18.7	19.1

C.2 MATH Prompt

Listing 2 shows the 4-shot prompt used when sampling answers to MATH questions. We picked it by choosing 8 random examples from MATH and selecting examples which did not include Asymptote plotting commands. We chose four examples so that most problems fit within a context length of 1024, to enable comparisons with a wide range of models.

```
Problem:
Find the domain of the expression  $\frac{\sqrt{x-2}}{\sqrt{5-x}}$.}

Solution:
The expressions inside each square root must be non-negative. Therefore,
$x-2 \ge 0$, so $x\ge2$, and $5 - x \ge 0$, so $x \le 5$. Also, the denominator
cannot be equal to zero, so $5-x>0$, which gives $x<5$. Therefore, the domain of
the expression is $\boxed{[2,5)}$.
Final Answer: The final answer is $[2,5)$. I hope it is correct.

Problem:
If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12,$ then find
$\det (\mathbf{A} \mathbf{B}).$

Solution:
We have that $\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B})
= (2)(12) = \boxed{24}.$
Final Answer: The final answer is $24$. I hope it is correct.

Problem:
Terrell usually lifts two 20-pound weights 12 times. If he uses two 15-pound
weights instead, how many times must Terrell lift them in order to lift the
same total weight?

Solution:
If Terrell lifts two 20-pound weights 12 times, he lifts a total of
$2\cdot 12\cdot20=480$ pounds of weight.  If he lifts two 15-pound
weights instead for $n$ times, he will lift a total of $2\cdot15\cdot n=30n$
pounds of weight.  Equating this to 480 pounds, we can solve for $n$:
\begin{align*}
30n&=480\\
\Rightarrow\qquad n&=480/30=\boxed{16}
\end{align*}
Final Answer: The final answer is $16$. I hope it is correct.

Problem:
If the system of equations

\begin{align*}
6x-4y&=a,\\
6y-9x &=b.
\end{align*}has a solution $(x, y)$ where $x$ and $y$ are both nonzero,
find $\frac{a}{b},$ assuming $b$ is nonzero.

Solution:
If we multiply the first equation by $-\frac{3}{2}$, we obtain

$$6y-9x=-\frac{3}{2}a.$$Since we also know that $6y-9x=b$, we have

$$-\frac{3}{2}a=b\Rightarrow\frac{a}{b}=\boxed{-\frac{2}{3}}.$$
Final Answer: The final answer is $-\frac{2}{3}$. I hope it is correct.
```

Listing 2: 4-shot prompt used for MATH problems.

D Additional Investigations

D.1 Dependence of performance on number of generated samples

We study the dependence of performance on the number of generated samples per question on MATH and GSM8k. Table 6 shows results for maj1@k and maj5@k, and Figure 5 shows the dependence on k for `pass@k` and majority voting. We observe that while `pass@k` continues to improve, majority voting saturates quickly.

Table 6: Performance on MATH ($k = 256$) and GSM8k ($k = 100$) when generating k samples per task.

	MATH	GSM8k
Minerva 8B, maj1@k	25.4%	28.4%
Minerva 8B, maj5@k	47.6%	56.8%
Minerva 62B, maj1@k	43.4%	67.5%
Minerva 62B, maj5@k	64.9%	89.0%
Published SOTA	6.9%	74.5%

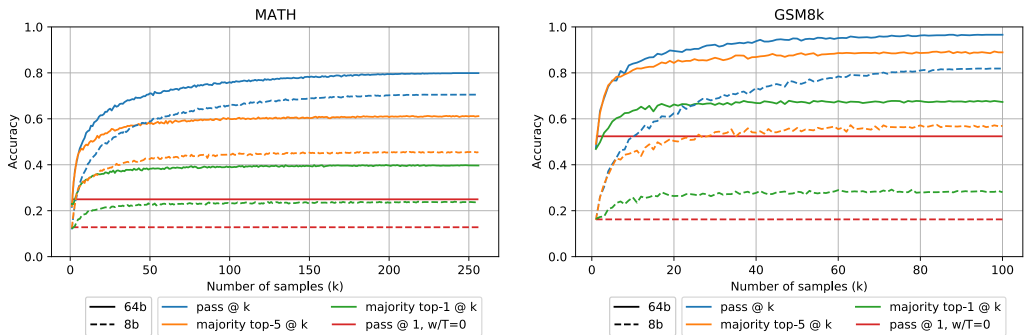

Figure 5: Accuracy as a function of k, the number of samples per task. Majority voting performance saturates quickly while `pass@k` seems to continue improving slowly. Accuracies were computed using exact string match (without `SymPy` processing).

D.2 Log-Likelihood Reranking

Table 7 compares majority voting with reranking based on the log-likelihood that the model assigns to each response. We observe that majority voting is significantly better.

Table 7: A comparison of the majority voting results presented in the main text with log-likelihood reranking. We do not use `SymPy` processing here.

	MATH
Minerva 62B, `pass1` $T = 0.0$	26.5%
Minerva 62B, Majority Voting 1@k	42.0%
Minerva 62B, `pass1` $T = 0.6$	21.8%
Minerva 62B, Log-likelihood 1@k	23.8%

D.3 Finetuning on MATH

Most of our results involve few-shot prompting Minerva on MATH and other datasets on which the model was not explicitly trained. In this section we discuss finetuning our models on the training split of the MATH dataset, and then evaluating on the test split as before. We finetune both the BaseModel and Minerva 8B for 3000 steps with 2048 tokens per batch with batch size 128 and dropout of 0.1. Similar to [17], we found that the accuracy for BaseModel kept improving despite the test loss increasing. We picked the model with the best test accuracy after 50 training steps.

We finetuned using a few different prompts: A 0-shot prompt, our custom 4-shot prompt, and a prompt containing 4 random examples. Each model was evaluated using the same prompt as was used during finetuning, except for the random prompt model, with was evaluated using the fixed 4-shot prompt that we used for the non-finetuned models.

The results can be found in Table 8. Standard finetuning does not seem to improve the performance of Minerva. On the other hand, it does lead to measurable improvements in BaseModel, though this performance still lagged behind Minerva. These results suggest that the marginal utility of supervised finetuning decreases as one improves the quality and diversity of the unsupervised pretraining or unsupervised finetuning dataset.

Table 8: We finetune BaseModel and Minerva using different finetuning methods. We find that while finetuning helps considerably for the BaseModel, it does not help for Minerva.

	MATH Accuracy	
	BaseModel 8B	Minerva 8B
Few Shot	1.5%	14.1%
Custom prompt finetuning	5.6 %	13.4%
Random prompt finetuning	4.4 %	12.9%
No prompt finetuning	5.6 %	13.0%

D.4 Majority Voting Thresholds

From Figure 5, we see how majority voting saturates rather quickly at some k, while `pass@k` keeps improving. Here we analyze the asymptotic behavior of majority voting at large k.

Let c_i denote the sorted number of counts for answer i when we sample N times and let there be a total for A_N answers. In other words, $\sum_{i=1}^{A_N} c_i = N, c_i > c_{i+1}$. We expect that when sampling $k \ll N$ samples, we can model the sampling distribution as a multinomial distribution with probabilities $p_i = \frac{c_i}{N}$. This approximation will have the error of attributing $p_i = 0$ to any answer which doesn't appear in N draws, so we can't really resolve probabilities smaller than $1/N$. This issue will not matter for our purposes as long as the maximum probability p_1 is significantly higher than $1/N$.

If we draw k samples from this multinomial distribution, we expect to *not* be able to identify the majority answer with 95% confidence as long as

$$p_1 - 2\sqrt{p_1(1-p_1)}/\sqrt{k} < p_2 + 2\sqrt{p_2(1-p_2)}/\sqrt{k} \tag{1}$$

$$k < \frac{4(\sqrt{p_1(1-p_1)} + \sqrt{p_2(1-p_2)})}{(p_1-p_2)^2} <_{p_1=p_2=0.5} \frac{4}{(p_1-p_2)^2} \tag{2}$$

For $k = 64$, this bound implies that the resolution for $p_1 - p_2$ is 0.25, but this is a very rough estimate. However, this exercise quantifies why and how majority voting saturates even if `pass@k` doesn't.

Another point is that in order to obtain the majority solution with 95% confidence, we need

$$p_1 - 2\sqrt{p_1(1-p_1)}/\sqrt{k} > 0 \rightarrow k > 4(1/p_1 - 1), p_1 > \frac{1}{k/4+1} \tag{3}$$

for $k = 64$, we can probe up to $p_1 > 0.06$.

E OCWCourses Evaluation Dataset Details

E.1 Breakdown of courses

Table 9 shows the breakdown of problems in our dataset by course. See Table 10 for a breakdown of problems by solution type.

E.2 Contractor instructions

Figure 6 shows the instructions provided to our contractor workforce.

Table 9: Problems of the OCWCourses dataset broken down by course.

Course	No. problems
Solid State Chemistry	97
Introduction to Astronomy	53
Differential equations	48
Dynamics and Control	26
Principles of Microeconomics	18
Special Relativity	11
Physical Chemistry	11
Ecology	5
Information and Entropy	3

Table 10: Answer types in OCWCourses

Answer Type	No. Problems
Numeric	191
Symbolic	81
Total	272

We would like to build a dataset of clean self-contained STEM problems and solutions written in clean and correct LaTeX code.
This dataset should have the following properties:

- **Self-contained problems with no external references**: A human should be able to solve each problem and understand the given solution without having to reference any other sources. For example, some problems reference lecture notes or a textbook. These problems should be rewritten to include the referenced information. If it takes you more than roughly five minutes to find the referenced material, please delete the problem; do not include it in the final submission.
- **No extraneous material**: The raw dataset contains extraneous data, such as headers, footers, problem numbers, and point values for problems. All of this data should be removed, so that each problem/solution pair contains only the content of the problem.
- **Clearly marked final answers**: For some problems, the solution ends in a specific value that constitutes the final answer. For example, a problem might ask the student to compute the value of an integral. In this case, the steps for computing the integral are part of the solution, but the expression that represents the antiderivative is the final answer (or in the case of a definite integral, the numerical value). When a problem has such a final answer, we ask that you annotate it using a special annotation. If such a final answer is not available, we ask that you try to define one yourself that represents the solution to the problem (though in some cases this will not be possible).
- **Including images and annotating non-essential images** If there are images in the problem, please include them with a single-line includegraphics command in the same way that they appear in the raw input files. To make the image render nicely, you can add a [scale=...] modifier; just make sure the command is on one line.

Figure 6: Instructions provided to contractors who worked on OCWCourses.

E.3 OCWCourses Prompt

```
Problem:
Subproblem 0: What is the net charge of arginine in a solution of $\mathrm{pH} 1.0$?
Please format your answer as +n or -n.
Solution:
The answer is +2.
Final answer: The final answer is +2. I hope it is correct.

Problem:
Subproblem 0: Let $z = 1 + \sqrt{3} i$. Find $a, b$ that satisfy the equation
$z^4 = a + bi$. Express your answer as the ordered pair $(a,b)$.
Solution:
$z^{4}$ has argument $4 \pi / 3$ and radius 16 , so it's equal to $-8-8 \sqrt{3} i$.
Thus $a = -8, b = -8\sqrt 3$, and our answer is $\boxed{(-8, -8\sqrt{3})}$.
Final answer: The final answer is (-8, -8\sqrt{3}). I hope it is correct.

Problem:
Preamble: For each Laplace Transform \(Y(s)\), find the function \(y(t)\):
Subproblem 0:
\[Y(s)=\boxed{\frac{1}{(s+a)(s+b)}}\]
Solution:
We can simplify with partial fractions:
\[Y(s)=\\frac{1}{(s+a)(s+b)}=\\frac{C}{s+a}+\\frac{D}{s+b}\]\nfind the constants
\(C\) and \(D\) by setting \(s=-a\) and \(s=-b\)
\[
  \begin{aligned}
  \frac{1}{(s+a)(s+b)} &=\\frac{C}{s+a}+\\frac{D}{s+b} \\\\
  1 &=C(s+b)+D(s+a) \\
  C &=\\frac{1}{b-a} \\
  D &=\\frac{1}{a-b}
  \end{aligned}
\]
therefore
\[\nY(s)=\frac{1}{b-a} \frac{1}{s+a}-\frac{1}{b-a} \frac{1}{s+b}
\]
By looking up the inverse Laplace Transform of \(\frac{1}{s+b}\), we find the total
solution \(y(t)\)
\[
  y(t)=\boxed{\frac{1}{b-a}\left(e^{-a t}-e^{-b t}\right)}
\].
Final answer: The final answer is \[\frac{1}{b-a}\left(e^{-a t}-e^{-b t}\right)\].
I hope it is correct.

Problem:
Preamble: The following subproblems refer to the differential equation
$\ddot{x}+b \dot{x}+x=0$.
Subproblem 0: What is the characteristic polynomial $p(s)$ of
$\ddot{x}+b \dot{x}+x=0$?
Solution:
The characteristic polynomial is $p(s)=\\boxed{s^{2}+b s+1}$.
Final answer: The final answer is $s^{2}+b s+1$. I hope it is correct.
```

Listing 3: Prompt used for OCWCourses.

E.4 Problems in OCWCourses

We provide the problems in OCWCourses as a separate file.

E.5 OCWCourses evaluation

As with the MATH dataset, special care must be taken in order to correctly extract answers and evaluate them for correctness. Here we describe how final answers are extracted and normalized. See Listing 4 for the code. During dataset creation, contractors annotated all automatically-verifiable solutions as belonging to one of several types: `symbolicexpression`, `symbolicequation`, or `numeric`. For `symbolicexpression` and `symbolicequation` answers, our approach is to convert the answer strings into `SymPy` quantities, and check equality programmatically. For numeric quantities, we first remove any units from the answer string, then convert the answer string to a float. If either numeric quantity is close to zero, our equality condition is that the absolute value of their difference is less than a threshold (0.01) of their mean; otherwise, we use the `numpy.isclose()` comparison.

As with MATH, we first extract the final answer from the full model response, which potentially includes chain-of-thought reasoning. In the few-shot prompt, we used the format `Final`

`Answer: The final answer is ANSWER. I hope it is correct.` for every final answer. We look for this pattern in the model output and extract `ANSWER`.

```python
import numpy as np

def get_answer(s: str) -> str:
  end_str = "I hope it is correct"
  start_str = "Final answer: "
  replacement_str = "The final answer is "

  scrub_periods = lambda x: x.strip().rstrip('.').strip()

  try:
    ans = s.split(end_str)[0].split(start_str)[1].strip().replace(replacement_str, "")
    ans = scrub_periods(ans)
    return ans
  except:
    print("Answer extraction failed")
    return None

def grade_question(question: dict) -> bool:
  """Grades a question."""
  formatting_fns = {'automatic:symbolicexpression': normalize_symbolic_expression,
                    'automatic:symbolicequation': normalize_symbolic_equation,
                    'automatic:numeric': normalize_numeric,
                    }
  question_type = question['type']
  formatting_fn = formatting_fns[question_type]

  # Get ground truth answer
  ground_truth_answer = formatting_fn(question['target'])
  if ground_truth_answer is None:
    raise ValueError("Could not parse question target answer")

  # Get model's answer
  model_answer = formatting_fn(get_answer(question['model_outputs'][0]))

  # Perform comparison
  grading_fns = {
                 'automatic:symbolicexpression': symbolic_equality,
                 'automatic:symbolicequation': lambda x,y: x == y,
                 'automatic:numeric': numeric_equality,
                 }

  return grading_fns[question_type](model_answer, ground_truth_answer)

def normalize_numeric(s):
  if s is None:
    return None
  for unit in ['eV',
               ' \\mathrm{~kg} \\cdot \\mathrm{m} / \\mathrm{s}',
               ' kg m/s', 'kg*m/s', 'kg', 'm/s', 'm / s', 'm s^{-1}',
               '\\text{ m/s}',
               ' \\mathrm{m/s}',
               ' \\text{ m/s}',
               'g/mole','g/mol',
               '\\mathrm{~g}',
               '\\mathrm{~g} / \\mathrm{mol}',
               'W',
               'erg/s',
               'years',
               'year',
               'cm']:
    s = s.replace(unit, '')
    s = s.strip()
  for maybe_unit in ['m', 's', 'cm']:
    s = s.replace('\\mathrm{'+maybe_unit+'}','')
    s = s.replace('\\mathrm{~'+maybe_unit+'}','')
    s = s.strip()
  s = s.strip('$')
  try:
    return float(eval(s))
  except:
    try:
      expr = parse_latex(s)
      if expr.is_number:
        return float(expr)
      return None
    except:
      return None

def numeric_equality(n1, n2, threshold=0.01):
  if n1 is None or n2 is None:
```

```python
    return False
  if np.isclose(n1,0) or np.isclose(n2,0) or np.isclose(n1-n2,0):
    return np.abs(n1-n2) < threshold * (n1+n2)/2
  else:
    return np.isclose(n1, n2)

def symbolic_equality(x,y):
  if x is None or y is None:
    return False
  else:
    try:
      return sympy.simplify(x-y) == 0
    except:
      return False

def normalize_symbolic_equation(s: Optional[str]):
  if not isinstance(s, str):
    return None
  if s.startswith('\\['):
    s = s[2:]
  if s.endswith('\\]'):
    s = s[:-2]
  s = s.replace('\\left(', '(')
  s = s.replace('\\right)', ')')
  s = s.replace('\\\\', '\\')
  if s.startswith('$') or s.endswith('$'):
    s = s.strip('$')
  try:
    maybe_expression = parse_latex(s)
    if not isinstance(maybe_expression, sympy.core.relational.Equality):
      # we have equation, not expression
      return None
    else:
      return maybe_expression
  except:
    return None

def normalize_symbolic_expression(s: Optional[str]):
  if not isinstance(s, str):
    return None
  if s.startswith('\\['):
    s = s[2:]
  if s.endswith('\\]'):
    s = s[:-2]
  s = s.replace('\\left(', '(')
  s = s.replace('\\right)', ')')
  s = s.replace('\\\\', '\\')
  if s.startswith('$') or s.endswith('$'):
    s = s.strip('$')
  try:
    maybe_expression = parse_latex(s)
    if isinstance(maybe_expression, sympy.core.relational.Equality):
      # we have equation, not expression
      return None
    if isinstance(maybe_expression, sympy.logic.boolalg.BooleanFalse):
      return None
    else:
      return maybe_expression
  except:
    return None
```

Listing 4: Python code used to normalize final answers.

F MMLU-STEM Evaluation Details

MMLU-STEM consists of the following 18 subtopics: `abstract_algebra`, `astronomy`, `college_biology`, `college_chemistry`, `college_computer_science`, `college_mathematics`, `college_physics`, `computer_security`, `conceptual_physics`, `electrical_engineering`, `elementary_mathematics`, `high_school_biology`, `high_school_chemistry`, `high_school_computer_science`, `high_school_mathematics`, `high_school_physics`, `high_school_statistics`, `machine_learning`.

The standard way of evaluating on MMLU is to construct a 5-shot prompt out of the dev set and then choose the option with the highest score [12]. This is what we report for `pass@1`.

We make use of the reasoning skills of the model and combine this task with chain of thought. To do this, we use a prompt which has a chain of thought before outputting the final answer. We extract the model answer by from model output of the form "`Final Answer: The final answer is CHOICE. I hope it is correct.`". When scoring choices, we use the real probability of each choice. In the chain-of-thought case, we can estimate the most probable choice (independently of the rationale) by picking the majority answer. Given that the set of possible final answers is reduced: $\texttt{CHOICE} \in \{A, B, C, D\}$ (as opposed to generative tasks where the set of possible answers was unbounded), we expect that we do not need many samples to find the majority option, and we therefore pick $k = 16$.

We use a multiple choice version of the MATH prompt (see Listing 5) for the subtopics which use equations: `abstract_algebra`, `college_mathematics`, `college_physics`, `elementary_mathematics`, `high_school_mathematics`, `high_school_physics`, `high_school_statistics`. We wrote a custom chain-of-thought for each of the remaining original prompts. Those prompts can be found in the supplementary materials.

```
Problem:
Find the domain of the expression $\frac{\sqrt{x-2}}{\sqrt{5-x}}$.
What of the following is the right choice? Explain you answer.
(A) [-5,-2), (B) [2,5), (C) [-2,-5), (D) [5,2)
Solution:
The expressions inside each square root must be non-negative. Therefore, $x-2 \ge 0$, so $x\ge2$, and $5 - x \ge 0$, so $x \le 5$. Also, the denominator cannot be equal to zero, so $5-x>0$, which gives $x<5$. Therefore, the domain of the expression is $\boxed{[2,5)}$.
Final Answer: The final answer is (B). I hope it is correct.

Problem:
If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12,$ then find $\det (\mathbf{A} \mathbf{B}).$
What of the following is the right choice? Explain you answer.
(A) 14, (B) 4, (C) 2, (D) 24
Solution:
We have that $\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}.$
Final Answer: The final answer is (D). I hope it is correct.

Problem:
Terrell usually lifts two 20-pound weights 12 times. If he uses two 15-pound weights instead, how many times must Terrell lift them in order to lift the same total weight?
What of the following is the right choice? Explain you answer.
(A) 12, (B) 20, (C) 16, (D) 15
Solution:
If Terrell lifts two 20-pound weights 12 times, he lifts a total of $2\cdot 12\cdot20=480$ pounds of weight.  If he lifts two 15-pound weights instead for $n$ times, he will lift a total of $2\cdot15\cdot n=30n$ pounds of weight.  Equating this to 480 pounds, we can solve for $n$: \begin{align*}
30n&=480\\
\Rightarrow\qquad n&=480/30=\boxed{16}
\end{align*}
Final Answer: The final answer is (C). I hope it is correct.

Problem:
If the system of equations

\begin{align*}
6x-4y&=a,\\
6y-9x &=b.
\end{align*}has a solution $(x, y)$ where $x$ and $y$ are both nonzero, find $\frac{a}{b},$ assuming $b$ is nonzero.
What of the following is the right choice? Explain you answer.
(A) $-\frac{2}{3}$, (B) $\frac{2}{3}$, (C) $\frac{1}{3}$, (D) $\frac{4}{9}$
Solution:
If we multiply the first equation by $-\frac{3}{2}$, we obtain

$$6y-9x=-\frac{3}{2}a.$$Since we also know that $6y-9x=b$, we have

$$-\frac{3}{2}a=b\Rightarrow\frac{a}{b}=\boxed{-\frac{2}{3}}.$$
Final Answer: The final answer is (A). I hope it is correct.
```

Listing 5: Multiple choice version of MATH prompt.

G Evaluation on natural-language inference task

In addition to the evaluation tasks mentioned in the main text, we also evaluate Minerva (540B) on a natural-language inference task known as EQUATE [50], in which quantitative reasoning is useful for determining the answer. We perform 4-shot prompting, using manually-designed prompt questions which were written to be similar to the questions in the dataset, but are augmented with chain-of-thought reasoning for each question; we do not do any majority voting for this evaluation, but sample a single model response at zero temperature.

An example prompt is shown below:

```
The following are examples of sentences which either entail or contradict one another.
Example
Sentence 1: Mary had 10 dollars yesterday; today, her father gave her 4 dollars.
Sentence 2: Mary has 14 dollars now.
Reasoning: According to sentence 1, Mary now has 10+4=14 dollars, as sentence 2 claims. So sentence 1 implies sentence 2, an example of entailment.
Label: Entailment

Example
Sentence 1: Mary had 10 dollars yesterday; today, her father gave her 7 dollars.
Sentence 2: Mary has 16 dollars now.
Reasoning: According to sentence 1, Mary now has 10+7=17 dollars; but sentence 2 claims differently, that she has 16 dollars.  Since the claim of sentence 2 directly contradicts that of sentence 1, this is an example of contradiction.
Label: Contradiction

Example
Sentence 1: A restaurant served 5 sandwiches during lunch and 3 sandwiches during dinner today .
Sentence 2: 8 sandwiches were served today
Reasoning: Using the numbers in sentence 1, we find that the restaurant served a total of 5 + 3 = 8 pizzas, which is the claim of sentence 2.  Thus, this pair of sentences demonstrates entailment.
Label: Entailment

Example
Sentence 1: A restaurant served 5 sandwiches during lunch and 6 during dinner today .
Sentence 2: 17 sandwiches were served today
Reasoning: The restaurant served a total of 5 + 6 = 11 sandwiches, but sentence 2 claims it served 17 sandwiches.  Since 15 is not equal to 17, this pair of sentences forms a contradiction.
Label: Contradiction
```

Listing 6: EQUATE AWP (Arithmetic Word Problem) Prompt.

Results of our evaluation, on all five subdivisions of EQUATE, are shown in the table below:

Split	AWP	NewsNLI	RedditNLI	NR ST	RTE-Q
Minerva performance	**83.4**	62.6	**66.8**	48.7	64.5
Previous SOTA	71.5	**72.8**	58.8	**63.3**	**68.1**

H Arithmetic Results

We compute the performance of Minerva on basic arithmetic by supplying the following prompt:

```
Q: What is $12 + 34$?
A: $46$
Q: What is $21 * 67$?
A: $1407$
Q: What is $98 - 17$?
A: $81$
Q: What is $84763 + 23079$?
A: $107842$
Q: What is $84763 + 23079$?
A: $107842$
Q: What is $313091 * 27$?
A: $8453457$
Q: What is ${x} {op} {y}$?
A:
```

Listing 7: Arithmetic prompt.

For each operation, and for each choice of number of digits, we sample 500 arithmetic questions at random. Figure 7 shows the model's performance.

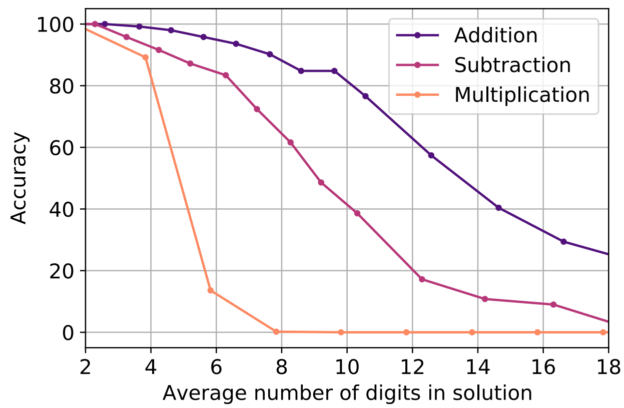

Figure 7: Accuracy on basic arithmetic questions.

I Additional Model Samples

I.1 Additional Correct Samples

This is an example of a question we wrote ourselves:

Question: What are the eigenvalues of a 2-by-2 matrix $M = \begin{pmatrix} 1 & 2 \\ -1 & 5 \end{pmatrix}$?

Model output:
The characteristic polynomial of M is $|M - \lambda I|$, where I is the 2-by-2 identity matrix. So the characteristic polynomial is $|M - \lambda I| = \begin{vmatrix} 1-\lambda & 2 \\ -1 & 5-\lambda \end{vmatrix} = (1-\lambda)(5-\lambda) + 2 = \lambda^2 - 6\lambda + 7$. This is a quadratic polynomial, so the eigenvalues are the solutions of $\lambda^2 - 6\lambda + 7 = 0$. We can solve this by completing the square:

$$\begin{aligned} \lambda^2 - 6\lambda + 7 &= 0 \\ \Rightarrow \quad \lambda^2 - 6\lambda + 9 - 9 + 7 &= 0 \\ \Rightarrow \quad (\lambda - 3)^2 - 2 &= 0 \\ \Rightarrow \quad (\lambda - 3)^2 &= 2 \\ \Rightarrow \quad \lambda - 3 &= \pm\sqrt{2} \\ \Rightarrow \quad \lambda &= 3 \pm \sqrt{2} \end{aligned}$$

So the eigenvalues are $3 + \sqrt{2}$ and $3 - \sqrt{2}$.

Additional examples from MATH that are correctly solved by the 62B model

Question: Let

$$f(n) = \begin{cases} n^3 + 2n - 1 & \text{if } n > 1, \\ n - 1 & \text{if } n \le 1. \end{cases}$$

Find $f(0) + f(1) + f(2)$.

Target: Since $0 \le 1$, we use the second case to find $f(0) = 0 - 1 = -1$. Since $1 \le 1$, we again use the second case to find $f(1) = 1 - 1 = 0$. Since $2 > 1$, we use the first case to find $f(2) = 2^3 + 2(2) - 1 = 11$. Therefore, $f(0) + f(1) + f(2) = -1 + 0 + 11 = \boxed{10}$.

Model output: We have $f(0) = 0 - 1 = -1$, $f(1) = 1 - 1 = 0$, and $f(2) = 2^3 + 2 \cdot 2 - 1 = 8 + 4 - 1 = 11$. So $f(0) + f(1) + f(2) = -1 + 0 + 11 = \boxed{10}$.

Question: Mary has 6 identical basil plants, and three different window sills she can put them on. How many ways are there for Mary to put the plants on the window sills?

Target: Since the plants are indistinguishable, we must only count the number of plants on each window sill. If all the plants are on one window sill, there are 3 ways to choose which window sill they are on. If 5 plants are on one window sill and the last is on another, there are $3! = 6$ ways to choose which plants go on which window sill. If 4 plants are on one window sill and the last two are on another, there are $3! = 6$ ways to choose which window sill they are on. If 4 plants are on one window sill and the last two are each on one of the other windows, there are 3 ways to choose which window the 4 plants are on. If 3 plants are on one window and the other 3 plants are all on another window, there are 3 ways to choose which window has no plants. If 3 plants are on one window, 2 plants on another window, and 1 plant on the last window, there are $3! = 6$ ways to choose which plants are on which windows. If 2 plants are on each window, there is only one way to arrange them.
In total, there are $3 + 6 + 6 + 3 + 3 + 6 + 1 = \boxed{28}$ ways to arrange the plants on the window sills.

Model output: We can use the Binomial Theorem to solve this problem. We have 6 plants and 3 sills, so we have $6 + 3 - 1 = 8$ total objects. Since we have 8 objects, we can use the Binomial Theorem to expand $(x + y)^8$. The coefficient of x^6y^2 will be the number of ways for Mary to put the plants on the sills. We can expand $(x + y)^8$ using the Binomial Theorem:

$$(x+y)^8 = \binom{8}{0}x^8 + \binom{8}{1}x^7y + \binom{8}{2}x^6y^2 + \binom{8}{3}x^5y^3 + \binom{8}{4}x^4y^4 + \binom{8}{5}x^3y^5 + \binom{8}{6}x^2y^6 + \binom{8}{7}xy^7 + \binom{8}{8}y^8.$$

Since we are only interested in the coefficient of x^6y^2, we can ignore all terms that do not have x^6 and y^2. Therefore, we are left with $\binom{8}{2}x^6y^2$. We can evaluate $\binom{8}{2}$ to get 28. Therefore, there are 28 ways for Mary to put the plants on the sills.

I.2 Categorizing Model Mistakes

We used the following to label model mistakes:

- **Incorrect reasoning:** there is a mistake in how different steps in the chain-of-thought are connected.
- **Incorrect calculation:** symbolic or numeric manipulation mistake.
- **Uses incorrect fact:** uses a fact or equation which is wrong.
- **Misunderstood question:** approach to problem is wrong.
- **Answer too short:** wrong answer without justification.
- **Hallucinated math objects:** fabricated mathematical concepts or symbols with no basis in the problem.

Table 11: Failure modes of the 8B Minerva model, out of 201 samples which the 62B model solved correctly and the 8B model did not.

Type of mistakes	Occurrences
Incorrect reasoning	82
Incorrect calculation	70
Misunderstands question	22
Uses incorrect fact	16
Solution too short	4
Hallucinated math objects	4
Other mistakes	3

We include 2 samples for each of the main failure categories. Notice that some categorization depended on human judgement and in some instances more than one label was applied.

Incorrect calculation.

Question: If $\sqrt{400} = \sqrt{81} + \sqrt{n}$, then what is the value of n?

Target: Not to be fooled by the square roots, we rewrite the equation as $20 = 9 + \sqrt{n}$. Thus, $\sqrt{n} = 11$ and $n = \boxed{121}$. Final Answer: The final answer is 121. I hope it is correct.

Model output (8B): We have that $\sqrt{400} = \sqrt{81} + \sqrt{n}$. Therefore, $400 = 81 + n$, which gives $319 = n$. Final Answer: The final answer is 319.

Model output (62B): We have $\sqrt{400} = \sqrt{81} + \sqrt{n} \Rightarrow \sqrt{16 \cdot 25} = \sqrt{9 \cdot 9} + \sqrt{n}$. Since $\sqrt{16 \cdot 25} = \sqrt{16} \cdot \sqrt{25} = 4 \cdot 5 = 20$ and $\sqrt{9 \cdot 9} = \sqrt{9} \cdot \sqrt{9} = 3 \cdot 3 = 9$, we have $20 = 9 + \sqrt{n}$. Subtracting 9 from both sides gives $11 = \sqrt{n}$. Squaring both sides gives $11^2 = n$, so $n = \boxed{121}$.
Final Answer: The final answer is 121.

Question: If $y = 8 - 5x + 4x^2$, what is the value of y when $x = -2$?

Target: We have $y = 8 - 5x + 4x^2 = 8 - 5(-2) + 4(-2)^2 = 8 + 10 + 4(4) = 8 + 10 + 16 = \boxed{34}$.
Final Answer: The final answer is 34. I hope it is correct.

Model output (8B): We have that $y = 8 - 5x + 4x^2$. When $x = -2$, we have that $y = 8 - 5(-2) + 4(-2)^2$, so $y = 8 + 10 - 16 = 4$. Final Answer: The final answer is 4.

Model output (62B): If $y = 8 - 5x + 4x^2$ and $x = -2$, then

$$y = 8 - 5(-2) + 4(-2)^2 = 8 + 10 + 16 = \boxed{34}.$$

Final Answer: The final answer is 34.

Incorrect reasoning.

Question: At each basketball practice last week, Jenny made twice as many free throws as she made at the previous practice. At her fifth practice she made 48 free throws. How many free throws did she make at the first practice?

Target: At Jenny's fourth practice she made $\frac{1}{2}(48) = 24$ free throws. At her third practice she made 12, at her second practice she made 6, and at her first practice she made $\boxed{3}$. Final Answer: The final answer is 3. I hope it is correct.

Model output (8B): At the first practice, Jenny made $\frac{48}{2} = 24$ free throws. Final Answer: The final answer is 24.

Model output (62B): Let n be the number of free throws that Jenny made at the first practice. We know that she made 48 free throws at the fifth practice. We also know that she made twice as many free throws at the ith practice as she made at the $(i-1)$th practice. Therefore, we have the following sequence of equations:

$$\begin{aligned} 48 &= 2 \cdot n \\ 48 &= 2 \cdot 2 \cdot n \\ 48 &= 2 \cdot 2 \cdot 2 \cdot n \\ &\vdots \\ 48 &= 2^{5-1} \cdot n = 2^4 \cdot n \\ \Rightarrow \qquad n &= 48 \cdot 2^{-4} = 48 \cdot 1/16 = \boxed{3} \end{aligned}$$

Final Answer: The final answer is 3.

We include below one additional example of a reasoning error committed by the 62B model and fixed by the 540B model. The project webpage features more examples comparing these two models.

Question: Find the sum of the base-2 geometric series $0.1_2 - 0.01_2 + 0.001_2 - 0.0001_2 + 0.00001_2 \ldots$; give your answer as a fraction in which the numerator and denominator are both expressed in base 10.

Target: Places to the right of the decimal point represent negative powers of the base, so we notice that the series in base 10 is $2^{-1} - 2^{-2} + 2^{-3} \ldots = \frac{1}{2} - \frac{1}{4} + \frac{1}{8} \ldots$. We recognize this as a geometric series with a common ratio of $-\frac{1}{2}$ and apply the formula $\frac{a}{1-r}$ for the sum of a geometric series with first term a and common ratio r. We get

$$\frac{\frac{1}{2}}{1 - \left(-\frac{1}{2}\right)} = \frac{\frac{1}{2}}{\frac{3}{2}} = \frac{1}{3}.$$

So, the sum of the geometric series is $\boxed{\frac{1}{3}}$. Final Answer: The final answer is $\frac{1}{3}$. I hope it is correct.

Model output (62B): The sum of the geometric series is $\frac{0.1_2}{1-0.1_2} = \frac{1/2}{1-1/2} = \frac{1/2}{1/2} = \boxed{1}$. Final Answer: The final answer is 1. I hope it is correct

Model output (540B): Since $0.1_2 = \frac{1}{2}$, $0.01_2 = \frac{1}{4}$, $0.001_2 = \frac{1}{8}$, $0.0001_2 = \frac{1}{16}$, and $0.00001_2 = \frac{1}{32}$, the given series is

$$\frac{1}{2} - \frac{1}{4} + \frac{1}{8} - \frac{1}{16} + \frac{1}{32} \ldots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{2^n}.$$

Since

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{2^n} = \frac{1}{2} \sum_{n=1}^{\infty} (-1)^{n+1} \left(\frac{1}{2}\right)^{n-1} = \frac{1}{2} \cdot \frac{1}{1 - \left(-\frac{1}{2}\right)} = \frac{1}{2} \cdot \frac{1}{1 + \frac{1}{2}} = \frac{1}{2} \cdot \frac{1}{\frac{3}{2}} = \frac{1}{2} \cdot \frac{2}{3} = \boxed{\frac{1}{3}}.$$

Final Answer: The final answer is $\frac{1}{3}$. I hope it is correct.

Misunderstanding of a problem statement.

This label was applied when a human rater thought that an attempted solution may be a part of a mathematically sound solution to a different problem.

Question: Which of the following points is farthest from the origin: $(1, 4)$, $(3, 5)$, $(-6, 0)$, $(-4, -2)$?

Target: We find the distance for all the points using the distance formula: For $(1, 4)$: $\sqrt{(1-0)^2 + (4-0)^2} = \sqrt{17}$ For $(3, 5)$: $\sqrt{(3-0)^2 + (5-0)^2} = \sqrt{34}$ For $(-6, 0)$: $\sqrt{(-6-0)^2 + (0-0)^2} = \sqrt{36}$ For $(-4, -2)$: $\sqrt{(-4-0)^2 + (-2-0)^2} = \sqrt{20}$ Therefore, the point farthest from the origin is $\boxed{(-6, 0)}$. Final Answer: The final answer is $(-6, 0)$. I hope it is correct.

Model output (8B): We can find the distance between any two points (x_1, y_1) and (x_2, y_2) using the formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Therefore, the distance between $(1, 4)$ and $(3, 5)$ is

$$d = \sqrt{(3-1)^2 + (5-4)^2} = \sqrt{4+1} = \sqrt{5}.$$

Similarly, the distance between $(3, 5)$ and $(-6, 0)$ is

$$d = \sqrt{(-6-3)^2 + (0-5)^2} = \sqrt{9+25} = \sqrt{34}.$$

The distance between $(-6, 0)$ and $(-4, -2)$ is

$$d = \sqrt{(-4+6)^2 + (-2-0)^2} = \sqrt{4+4} = \sqrt{8} = \boxed{2\sqrt{2}}.$$

Since $2\sqrt{2}$ is the largest distance, $(-4, -2)$ is the farthest from the origin. Final Answer: The final answer is $(-4, -2)$.

Model output (62B): The distance of a point (x, y) from the origin is $d = \sqrt{x^2 + y^2}$. To find which point is farthest from the origin, we must find the largest value of d. To do this, we must find the largest value of $x^2 + y^2$.

$$1^2 + 4^2 = 1 + 16 = 17$$
$$3^2 + 5^2 = 9 + 25 = 34$$
$$(-6)^2 + 0^2 = 36 + 0 = 36$$
$$(-4)^2 + (-2)^2 = 16 + 4 = 20$$

Therefore, the largest value of $x^2 + y^2$ is 36, which means that $(-6, 0)$ is the farthest point from the origin. Final Answer: The final answer is $(-6, 0)$.

Question: John is 31 years younger than his dad. The sum of their ages is 53 years. How many years old is John's dad?

Target: Let j be John's age and d be his dad's age. We are trying to find the value of d. We can create a system of two equations to represent the given information. They are

$$j = d - 31$$
$$j + d = 53$$

We want to find d, so we need to eliminate j from the equations above. Substituting the first equation into the second to eliminate j, we have $(d - 31) + d = 53$, or $d = 42$. Thus, John's dad is $\boxed{42}$ years old. Final Answer: The final answer is 42. I hope it is correct.

Model output (8B): If John is x years old, his dad is $x + 31$ years old. The sum of their ages is $x + 31 + x = 53$. Therefore,

$$2x + 31 = 53 \Rightarrow 2x = 53 - 31 = 22 \Rightarrow x = \frac{22}{2} = \boxed{11}.$$

Final Answer: The final answer is 11.

Model output (62B): Let x denote John's age and y denote his dad's age. Since John is 31 years younger than his dad, we have $x = y - 31$. Since the sum of their ages is 53 years, we have $x + y = 53$. Substituting $y - 31$ for x into the second equation, we have

$$\begin{aligned} y - 31 + y &= 53 \\ \Rightarrow \qquad 2y &= 53 + 31 = 84 \\ \Rightarrow \qquad y &= \frac{84}{2} = \boxed{42}. \end{aligned}$$

Final Answer: The final answer is 42.

Uses of incorrect facts.

We decided to separate this specific kind of a reasoning error — the model provides a consistent reasoning that relies on an additional fact that does not hold in the given context.

Question: If $f(x)$ is a polynomial of degree 3, and $g(x)$ is a polynomial of degree 5, then what is the degree of polynomial $2f(x) + 4g(x)$?

Target: Let $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ and $g(x) = b_5x^5 + b_4x^4 + b_3x^3 + b_2x^2 + b_1x + b_0$. Then

$$\begin{aligned} 2f(x) + 4g(x) &= 2(a_3x^3 + a_2x^2 + a_1x + a_0) \\ &\quad + 4(b_5x^5 + b_4x^4 + b_3x^3 + b_2x^2 + b_1x + b_0) \\ &= 4b_5x^5 + 4b_4x^4 + (2a_3 + 4b_3)x^3 + (2a_2 + 4b_2)x^2 \\ &\quad + (2a_1 + 4b_1)x + (2a_0 + 4b_0). \end{aligned}$$

Thus, the degree of $2f(x) + 4g(x)$ is $\boxed{5}$. Final Answer: The final answer is 5. I hope it is correct.

Model output (8B): The degree of $f(x)$ is 3, and the degree of $g(x)$ is 5, so the degree of $2f(x)+4g(x)$ is 8. Final Answer: The final answer is 8.

Model output (62B): If $f(x)$ is a polynomial of degree 3 and $g(x)$ is a polynomial of degree 5, then $2f(x) + 4g(x)$ is a polynomial of degree 5. Final Answer: The final answer is 5.

Question: Lines $y = (3a + 2)x - 2$ and $2y = (a - 4)x + 2$ are parallel. What is the value of a?

Target: We find the slopes of the two lines and set them equal to each other, since parallel lines have the same slope. This gives $3a + 2 = \frac{a}{2} - 2$, which implies $a = \boxed{-\frac{8}{5}}$. Final Answer: The final answer is $-\frac{8}{5}$. I hope it is correct.

Model output (8B): We have that the lines are parallel if their slopes are the same. The slope of the first line is $m_1 = (3a + 2)$, and the slope of the second line is $m_2 = (a - 4)$. The lines are parallel if $m_1 = m_2$, so $3a + 2 = a - 4$, and therefore $2a = -6$, so $a = -3$. Final Answer: The final answer is -3.

Model output (62B): If lines $y = (3a + 2)x - 2$ and $2y = (a - 4)x + 2$ are parallel, then their slopes are equal. Thus,

$$\frac{3a + 2}{1} = \frac{a - 4}{2} \Rightarrow 6a + 4 = a - 4 \Rightarrow 5a = -8 \Rightarrow \boxed{a = -\frac{8}{5}}.$$

Final Answer: The final answer is $-\frac{8}{5}$.

Hallucination.

This label was applied when a human rater discovered in a solution a mathematical object that cannot be easily tracked to original formulation of the task.

Question: What is the value of y in the arithmetic sequence $y + 6$, 12, y?

Target: The difference between the second and first term is $12 - (y + 6) = 6 - y$, and the difference between the third and second term is $y - 12$. These must be equal, so $6 - y = y - 12$. Solving for y, we find $y = \boxed{9}$. Final Answer: The final answer is 9. I hope it is correct.

Model output (8B): Since the sequence is arithmetic, the common difference is $12 - (y + 6) = y - 6$. Since the sequence starts at $y + 6$, we have $y + 6 = a_1 = 12$, and so $y = 6$. Final Answer: The final answer is 6.

Model output (62B): Since the sequence is arithmetic, there is a common difference, which we will call d. Since the second term is 12, and the first term is $y + 6$, we have $d = 12 - (y + 6) = 6 - y$. Since the third term is y, we have $d = y - 12$. Thus $6 - y = y - 12$, so $y = 9$.
Final Answer: The final answer is 9.

Question: What is the value of $\sqrt[3]{4^5 + 4^5 + 4^5 + 4^5}$?

Target: Simplify under the radical first: $4^5 + 4^5 + 4^5 + 4^5 = 4 \cdot 4^5 = 4^6$, and the cube root of 4^6 is $4^{6/3} = 4^2 = \boxed{16}$. Final Answer: The final answer is 16. I hope it is correct.

Model output (8B): The expression $\sqrt[3]{4^5 + 4^5 + 4^5 + 4^5}$ is equal to $\sqrt[3]{4^5} \cdot \sqrt[3]{1}$, which is equal to $4 \cdot \sqrt[3]{1} = \boxed{4}$. Final Answer: The final answer is 4.

Model output (62B): We have

$$\sqrt[3]{4^5 + 4^5 + 4^5 + 4^5} = \sqrt[3]{4^5 \cdot 4} = \sqrt[3]{4^6} = 4^{6/3} = \boxed{4^2}.$$

Final Answer: The final answer is 16.

Short and wrong solutions.

This label was applied when a human rater decided that there are missing key reasoning steps.

Question: Simplify and write the result with a rational denominator:

$$\sqrt{\sqrt[3]{\sqrt{\frac{1}{729}}}}$$

Target: First, notice that $729 = 3^6$. We can begin simplifying from the innermost square root:

$$\sqrt{\sqrt[3]{\frac{1}{\sqrt{729}}}} = \sqrt{\sqrt[3]{\frac{1}{27}}} = \sqrt{\frac{1}{3}} = \frac{1}{\sqrt{3}} = \boxed{\frac{\sqrt{3}}{3}}$$

Final Answer: The final answer is $\frac{\sqrt{3}}{3}$. I hope it is correct.

Model output (8B): The expression can be written as

$$\sqrt[3]{\sqrt{\frac{1}{729}}} = \sqrt[6]{729^{-1/2}} = \sqrt[6]{3^{-3}} = \boxed{\frac{1}{3}}.$$

Final Answer: The final answer is $\frac{1}{3}$.

Model output (62B): We have that

$$\begin{aligned}
\sqrt{\sqrt[3]{\sqrt{\frac{1}{729}}}} &= \sqrt{\sqrt[3]{\sqrt{\frac{1}{3^6}}}} \\
&= \sqrt{\sqrt[3]{\sqrt{\frac{1}{3^3 \cdot 3^3}}}} \\
&= \sqrt{\sqrt[3]{\sqrt{\frac{1}{27 \cdot 27}}}} \\
&= \sqrt{\sqrt[3]{\sqrt{\frac{1}{3^3 \cdot 3^3}}}} \\
&= \sqrt{\sqrt[3]{\frac{1}{3^3}}} \\
&= \sqrt{\frac{1}{3}} \\
&= \frac{\sqrt{3}}{3}
\end{aligned}$$

Final Answer: The final answer is $\frac{\sqrt{3}}{3}$.

I.3 False Positive for pass@k

Here we present a false positive sample analysis for the pass@k metric to complement the pass@1 analysis from the main text. The goal is to look for the hardest correct samples because that is most likely where false positives are the most prevalent.

In this way, we will focus on tasks which are labeled as correct by the pass@k metric but are not the majority answer. Among these we will pick the correct sample whose final answer only appears once.

For the 62B model, $k = 256$ samples, 2,000 tasks are solved by pass@k and not by majority voting $1@256$. Of these correct samples, we focus on the 270 which appear only once.

We manually looked into 25 of these samples and labeled them in one of five categories:

- False Positive with Graph (8%): false positive, required parsing an Asymptote graph
- Clear False Positive (16%): the model is producing the right answer using the wrong approach/method.
- False Positives with minor mistakes (16%): models have the right reasoning but make a minor mistake in reasoning. It is not clear if the model makes two mistakes that cancel or whether the model ignores the mistake in the reasoning and attends to parts of the reasonign that were correct.
- Correct Answers without Explanining Steps (16%): the reasoning is correct but sometimes the model plugs formulas too quickly and without explaning. This would remove some points in a proper grading.
- Correct Answers (44%): model gets it right.

The density of false positives in this set of samples is roughly 30%. There are $84\% - 43\% = 41\%$ tasks which are solved by pass@k and not by maj$1@k$. If we extrapolate the False Positive Rate (FPR), we expect the actual true positive samples to be $41\% \cdot 0.7 = 29\%$. If the FPR for the majority voting samples is 8%, we expect the accuracy to be at least $43\% \cdot 0.92 + 29\% = 65\%$. Note that to estimate the FPR for the samples which are only solved by pass@k , we looked for the location where there is less confidence about the answer: problems which only had 1 correct answer (as opposed to many correct final answers) and 90% of all correct tasks have more than one sample with the correct final answer.

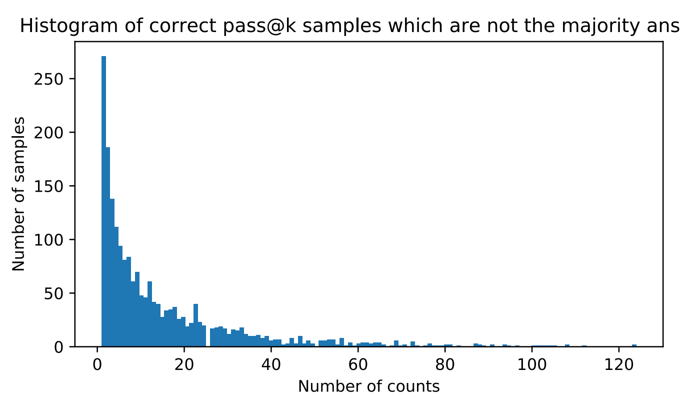

Figure 8: Distribution of "hardest" correct samples with respect to number of counts.

Examples of Correct Answers without Explaining Steps

Question: Evaluate

$$\sin(\arcsin 0.4 + \arcsin 0.5) \cdot \sin(\arcsin 0.5 - \arcsin 0.4).$$

Target: From the angle addition and subtraction formulas,

$$\begin{aligned}\sin(x+y) &= \sin x \cos y + \cos x \sin y, \\ \sin(x-y) &= \sin x \cos y - \cos x \sin y,\end{aligned}$$

so

$$\begin{aligned}\sin(x+y)\sin(x-y) &= (\sin x \cos y + \cos x \sin y)(\sin x \cos y - \cos x \sin y) \\ &= \sin^2 x \cos^2 y + \sin x \cos x \sin y \cos y - \sin x \cos x \sin y \cos y - \cos^2 x \sin^2 y \\ &= \sin^2 x (1 - \sin^2 y) - (1 - \sin^2 x) \sin^2 y \\ &= \sin^2 x - \sin^2 x \sin^2 y - \sin^2 y + \sin^2 x \sin^2 y \\ &= \sin^2 x - \sin^2 y.\end{aligned}$$

Taking $x = \arcsin 0.5$ and $y = \arcsin 0.4$, we get

$$\begin{aligned}\sin(\arcsin 0.5 + \arcsin 0.4) \cdot \sin(\arcsin 0.5 - \arcsin 0.4) &= \sin^2(\arcsin 0.5) - \sin^2(\arcsin 0.4) \\ &= 0.5^2 - 0.4^2 \\ &= 0.09 = \boxed{\frac{9}{100}}.\end{aligned}$$

Final Answer: The final answer is $\frac{9}{100}$. I hope it is correct.

Model output: The expression is equal to $\sin(\arcsin 0.4 + \arcsin 0.5) \cdot \sin(\arcsin 0.5 - \arcsin 0.4) = \sin(\arcsin(\frac{4}{10} + \frac{5}{10})) \cdot \sin(\arcsin(\frac{5}{10} - \frac{4}{10})) = \sin(\arcsin(\frac{9}{10})) \cdot \sin(\arcsin(\frac{1}{10})) = \frac{9}{10} \cdot \frac{1}{10} = \boxed{\frac{9}{100}}$.
Final Answer: The final answer is $\frac{9}{100}$. I hope it is correct.

Question: Find the product $10 \cdot \frac{4}{3} \cdot \frac{6}{4} \cdot \frac{8}{5} \cdot \frac{10}{6} \cdot \frac{12}{7} \cdot \frac{14}{8} \cdot \frac{16}{9} \cdot \frac{18}{10}$.

Target: The product of the numerators is

$$\begin{aligned}&(2 \cdot 2) \cdot (2 \cdot 3) \cdot (2 \cdot 4) \cdot (2 \cdot 5) \cdot (2 \cdot 6) \cdot (2 \cdot 7) \cdot (2 \cdot 8) \cdot (2 \cdot 9) \\ &= 2^8 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9.\end{aligned}$$

Then the given expression is equal to

$$10 \cdot \frac{2^8 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9}{3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10} = 2^9 = \boxed{512}.$$

Final Answer: The final answer is 512. I hope it is correct.

Model output: $10 \cdot \frac{4}{3} \cdot \frac{6}{4} \cdot \frac{8}{5} \cdot \frac{10}{6} \cdot \frac{12}{7} \cdot \frac{14}{8} \cdot \frac{16}{9} \cdot \frac{18}{10} = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^9 = \boxed{512}$. Final Answer: The final answer is 512. I hope it is correct.

Examples of False Positives with minor mistakes

Question:
The first 15 numbers of an arrangement are shown below. What would be the value of the 30^{th} number if the arrangement were continued?

$$\begin{aligned} &\text{Row } 1: 1 \\ &\text{Row } 2: 2, 2 \\ &\text{Row } 3: 3, 3, 3 \\ &\text{Row } 4: 4, 4, 4, 4 \\ &\text{Row } 5: 5, 5, 5, 5, 5 \end{aligned}$$

Target: Since we are told there are 15 numbers in the first 5 Rows, we want to find the 15^{th} number starting with the first number in Row 6. Since there are 6 numbers in Row 6, there are 7 numbers in Row 7 and there are 8 numbers in Row 8, the 15^{th} number if we start counting in Row 6 is located at the 2^{nd} spot of Row 8, which is of course an $\boxed{8}$. Final Answer: The final answer is 8. I hope it is correct.

Model output: The first 15 numbers of the arrangement are shown below.

$$\begin{aligned} &\text{Row } 1: 1 \\ &\text{Row } 2: 2, 2 \\ &\text{Row } 3: 3, 3, 3 \\ &\text{Row } 4: 4, 4, 4, 4 \\ &\text{Row } 5: 5, 5, 5, 5, 5 \end{aligned}$$

The 16^{th} number is 6. The 17^{th} number is 6. The 18^{th} number is 6. The 19^{th} number is 6. The 20^{th} number is 6. The 21^{st} number is 7. The 22^{nd} number is 7. The 23^{rd} number is 7. The 24^{th} number is 7. The 25^{th} number is 7. The 26^{th} number is 7. The 27^{th} number is 8. The 28^{th} number is 8. The 29^{th} number is 8. The 30^{th} number is 8.
Final Answer: The final answer is 8. I hope it is correct.

Question: If two positive numbers have arithmetic mean 2700 and harmonic mean 75, what is their geometric mean?

Target: Let the two numbers be a and b. If they have arithmetic mean 2700, then

$$\frac{a+b}{2} = 2700,$$

which gives us $a + b = 5400$. Since their harmonic mean is 75 we have

$$\frac{2}{\frac{1}{a}+\frac{1}{b}} = 75.$$

We can rearrange terms to get

$$\frac{1}{a}+\frac{1}{b} = \frac{2}{75}.$$

Taking the common denominator gives us

$$\frac{a+b}{ab} = \frac{2}{75}.$$

Substituting the value of $a + b$ and solving for ab gives

$$ab = \frac{5400 \cdot 75}{2} = 2700 \cdot 75.$$

Then the geometric mean is

$$\sqrt{ab} = \sqrt{2700 \cdot 75} = \boxed{450}.$$

Final Answer: The final answer is 450. I hope it is correct.

Model output: Let a and b be the two numbers. Then $\frac{a+b}{2} = 2700$ and $\frac{2ab}{a+b} = 75$. Dividing the second equation by the first gives $\frac{2ab}{a+b} = 75$. Then,

$$\begin{aligned}
& \frac{a+b}{2} = 2700 \\
\Rightarrow \quad & a+b = 5400 \\
\Rightarrow \quad & \frac{2ab}{a+b} = 75 \\
\Rightarrow \quad & \frac{2ab}{5400} = 75 \\
\Rightarrow \quad & \frac{ab}{2700} = 75 \\
\Rightarrow \quad & ab = 2700 \cdot 75 \\
\Rightarrow \quad & \sqrt{ab} = \sqrt{2700 \cdot 75} \\
\Rightarrow \quad & \sqrt{ab} = \sqrt{(30 \cdot 30)(75)} \\
\Rightarrow \quad & \sqrt{ab} = \sqrt{(30 \cdot 30)(5 \cdot 15)} \\
\Rightarrow \quad & \sqrt{ab} = 30 \cdot 15 \\
\Rightarrow \quad & \sqrt{ab} = \boxed{450}.
\end{aligned}$$

Final Answer: The final answer is 450. I hope it is correct.

Clear False Positive Examples

Question: How many integers in the list 25, 26, 27, ..., 250 are not perfect cubes?

Target: There are $250 - 25 + 1 = 226$ numbers in the list $25, 26, \ldots, 250$. We can find four perfect cubes in the list, namely $3^3, \ldots, 6^3$. So the number of non-perfect-cubes in the list is $226 - 4 = \boxed{222}$. Final Answer: The final answer is 222. I hope it is correct.

Model output: The first perfect cube in the list is $25^3 = 15625$. The last perfect cube is $250^3 = 15625000$. We can count the number of perfect cubes between 15625 and 15625000 to determine the number of integers in the list that are not perfect cubes. Since $\log_{10} 15625 = 4$ and $\log_{10} 15625000 = 7$, there are $7-4+1 = 4$ perfect cubes between 15625 and 15625000. Since the list has $250-25+1 = 226$ integers in total, there are $226 - 4 = 222$ integers in the list that are not perfect cubes. Final Answer: The final answer is 222. I hope it is correct.

Question: Arnold must water his cactus plants every 9 days and he must feed them every 60 days. He watered and fed them last Tuesday. What day of the week will it be the next time he waters and feeds them on the same day?

Target: The next time he feeds and waters them on the same day is after $\text{lcm}(9, 60) = 180$ days. Now to determine the day of the week we compute

$$180 \equiv 5 \pmod 7.$$

This day falls 5 days later in the week than a Tuesday, so he next feeds them on the same day on a $\boxed{\text{Sunday}}$. Final Answer: The final answer is Sunday. I hope it is correct.

Model output: Arnold must water his cactus plants every 9 days and feed them every 60 days. The least common multiple of 9 and 60 is $9 \cdot 2 \cdot 5 = 90$. Therefore, Arnold will water and feed them on the same day every 90 days. The next time he will do this is in $90 - 7 = 83$ days. Since there are 7 days in a week, this will be 12 days after the next Tuesday. So, the next time Arnold will water and feed his cactus plants on the same day will be $\boxed{\text{Sunday}}$. Final Answer: The final answer is Sunday. I hope it is correct.

Example False Positive with Graph

Question: In the diagram, what is the value of y? [asy] draw((5,0)–(0,0)–(0,5)); draw((.5,0)–(.5,.5)–(0,.5)); draw(Arc((0,0),1,90,360),Arrows); label("y°",(-1,-1),SW); [/asy]

Target: When the unknown angle y° is added to the 90° angle, the result is a complete rotation, or 360°. Thus, $y^\circ + 90^\circ = 360^\circ$ or $y = 360 - 90 = \boxed{270}$. Final Answer: The final answer is 270. I hope it is correct.

Model output: The angle in the diagram is the sum of the angle of the sector and the angle of the triangle. The angle of the sector is $360^\circ - 90^\circ = 270^\circ$. The angle of the triangle is $180^\circ - 90^\circ - 90^\circ = 0^\circ$. Therefore, the angle in the diagram is $270^\circ + 0^\circ = \boxed{270^\circ}$. Final Answer: The final answer is 270°. I hope it is correct.

I.4 Additional model samples

I.4.1 Polish National exam in Mathematics 2022

As conducted in May 2022, the actual exam contains 29 multiple-choice problems that we loosely frame as text problems, expecting the model to provide reasoning followed by an answer. We decided to skip in this evaluation three problems that rely on graphs.

Moreover, the exam contains six text problems where the student is expected to provide reasoning and a solution. In our evaluation, we included all these problems.

The model excels at more straightforward problems, making few reasoning or calculation mistakes. In particular, the model solves correctly 19 out of 26 multiple-choice problems. Mathematically valid reasoning backs the prevailing majority of the 19 correct solutions. Below we highlight selected positive examples as well as selected failure cases.

A more complete evaluation should include in-context learning based on previous iterations of the National Exam and assessment of other iterations of the exam. We leave this experiment as future work.

Question: Decide whether $(2\sqrt{8} - 3\sqrt{2})^2$ is 2, 1, 26 or 14.

Model output: $(2\sqrt{8} - 3\sqrt{2})^2 = (2\sqrt{2} \cdot 2\sqrt{2} - 3\sqrt{2})^2 = (4\sqrt{2} - 3\sqrt{2})^2 = (\sqrt{2})^2 = \boxed{2}$. Final solution: The final answer is 2.

Question: Positive numbers x and y satisfy the condition $2x = 3y$. Find the value of the expression $\frac{x^2+y^2}{x \cdot y}$.

Model output: $2x = 3y$, so $x = \frac{3}{2}y$. Substituting this into the expression, we have

$$\frac{x^2+y^2}{xy} = \frac{(\frac{3}{2}y)^2 + y^2}{\frac{3}{2}y^2} = \frac{\frac{9}{4}y^2 + y^2}{\frac{3}{2}y^2} = \frac{\frac{13}{4}y^2}{\frac{3}{2}y^2} = \frac{13}{4} \cdot \frac{2}{3} = \frac{13}{6}.$$

Final solution: The final answer is $\frac{13}{6}$.

Question: The price of the plot after two consecutive reductions, each time by 10% in relation to the price in force at the moment, is equal to USD 78732. Decide whether the price before reductions, rounded to 1 USD, is equal to USD 98732, USD 97200, USD 95266 or USD 94478.

Model output: Let x be the price before the reductions. Then, after the first reduction, the price is $0.9x$. After the second reduction, the price is $0.9(0.9x) = 0.81x$. Therefore, $0.81x = 78732$, so $x = 78732/0.81 = 97200$. The price before the reductions is USD $\boxed{97200}$. Final solution: The final answer is 97200.

Question: a_n is an geometric sequence so that $a_n > 0$ and $9 \cdot a_5 = 4 \cdot a_3$. Decide whether a_{n+1}/a_n is equal to $\frac{2}{3}$, $\frac{3}{2}$, $\frac{2}{9}$ or $\frac{9}{2}$.

Model output: $a_n = a_1 \cdot r^{n-1}$, so $a_5 = a_1 \cdot r^4$ and $a_3 = a_1 \cdot r^2$. Thus, $9 \cdot a_5 = 4 \cdot a_3$ implies that $9 \cdot a_1 \cdot r^4 = 4 \cdot a_1 \cdot r^2$. Dividing both sides by a_1 and r^2, we obtain $9 \cdot r^2 = 4$. Therefore, $r^2 = \frac{4}{9}$ and $r = \pm\frac{2}{3}$. Since $a_n > 0$, we must have $r = \frac{2}{3}$. Therefore, $a_{n+1}/a_n = \frac{a_1 \cdot r^n}{a_1 \cdot r^{n-1}} = r = \boxed{\frac{2}{3}}$. Final solution: The final answer is $\frac{2}{3}$.

Question: A parallelogram has sides of length 6 and 10 and the angle between them is 120°. Decide whether the area of the parallelogram is $30\sqrt{3}$, 30, $60\sqrt{3}$ or 60.

Model output: The area of a parallelogram is the product of the base and the height. The base is 6 and the height is $10 \sin 60^\circ = 10\sqrt{3}/2 = 5\sqrt{3}$. Therefore, the area is $6 \cdot 5\sqrt{3} = \boxed{30\sqrt{3}}$. Final solution: The final answer is $30\sqrt{3}$.

Question: Decide whether the number of different four-digit integers divisible by 5 and not divisible by 2 is $9 \cdot 8 \cdot 7 \cdot 2$, $9 \cdot 10 \cdot 10 \cdot 1$, $9 \cdot 10 \cdot 10 \cdot 2$ or $9 \cdot 9 \cdot 8 \cdot 1$.

Model output: The number of four-digit integers divisible by 5 is $9 \cdot 10 \cdot 10 \cdot 2$. The number of four-digit integers divisible by 10 is $9 \cdot 10 \cdot 10 \cdot 1$. Therefore, the number of four-digit integers divisible by 5 but not by 10 is $9 \cdot 10 \cdot 10 \cdot 2 - 9 \cdot 10 \cdot 10 \cdot 1 = 9 \cdot 10 \cdot 10 \cdot 1 = \boxed{900}$. Final solution: The final answer is 900.

An example of a calculation error of the 62B model fixed by the 540B model.

Question: For the arithmetic sequence a_n holds $a_1 = -1$ and $a_4 = 8$. Compute

$$\sum_{n=1}^{100} a_n.$$

Model output - 62B: We have $a_n = a_1 + (n-1)d$, where d is the common difference. So $a_4 = a_1 + 3d$ and $8 = -1 + 3d$, so $d = 3$. Now $a_n = -1 + 3(n-1) = 3n - 4$, so $\sum_{n=1}^{100} a_n = \sum_{n=1}^{100} (3n-4) = 3\sum_{n=1}^{100} n - 4\sum_{n=1}^{100} 1 = 3\frac{100 \cdot 101}{2} - 400 = 150 \cdot 101 - 400 = \boxed{15150}$. Final solution: The final answer is 15150.

Model output - 540B: Since we know $a_1 = -1$ and $a_4 = 8$, we can compute the difference d as $d = \frac{a_4 - a_1}{4-1} = \frac{9}{3} = 3$. Using the formula for arithmetic sequence, we have $a_{100} = a_1 + (100-1)d = -1 + 99 \cdot 3 = 296$. Now, we can evaluate our sum as follows:

$$\sum_{n=1}^{100} a_n = \frac{a_1 + a_{100}}{2} \cdot 100 = \frac{-1 + 296}{2} \cdot 100 = \boxed{14750}.$$

Final solution: The final answer is 14750.

An example of a reasoning error corrected by the 540B model.

Question: From $1, 2, 3, 4, 5, 6, 7, 8, 9$ we randomly draw with replacement numbers n and m. Compute the probability that $n \times m = 24$.
Model output - 62B: There are $9^2 = 81$ possible pairs of numbers. The only pair that satisfies $n \times m = 24$ is $(6, 4)$. Therefore, the probability is $\frac{1}{81}$. Final solution: The final answer is $\frac{1}{81}$.

Model output - 540B: We will have $n \times m = 24$ if and only if $n = 3$ and $m = 8$, or if $n = 4$ and $m = 6$, or if $n = 6$ and $m = 4$, or if $n = 8$ and $m = 3$. This is four cases out of the total of $9 \times 9 = 81$, so the probability is $\boxed{\frac{4}{81}}$. Final solution: The final answer is $\frac{4}{81}$

Question: From $1, 2, 3, 4, 5, 6, 7, 8, 9$ we randomly draw with replacement numbers n and m. Compute the probability that $n \times m = 24$.

Medium model [incorrect] There are $9^2 = 81$ possible pairs of numbers. The only pair that satisfies $n \times m = 24$ is $(6, 4)$. Therefore, the probability is $\frac{1}{81}$. Final solution: The final answer is $\frac{1}{81}$.
Large model [correct]: We will have $n \times m = 24$ if and only if $n = 3$ and $m = 8$, or if $n = 4$ and $m = 6$, or if $n = 6$ and $m = 4$, or if $n = 8$ and $m = 3$. This is four cases out of the total of $9 \times 9 = 81$, so the probability is $\boxed{\frac{4}{81}}$. Final solution: The final answer is $\frac{4}{81}$

An example of an error corrected by the 540B model.

Question: Points $A = (-2, 6)$ and $B = (3, b)$ are on a line that passes through $0, 0$. Decide whether b is 9, -9, -4 or 4.

Model output - 62B: The slope of the line is $\frac{b-6}{3+2} = \frac{b-6}{5}$. Since the line passes through the origin, the slope is $\frac{b-6}{5} = \frac{0-0}{0+0} = 0$. Therefore, $b - 6 = 0$ and $b = \boxed{6}$. Final solution: The final answer is 6.

Model output - 540B: Since A and B are on a line that passes through the origin, they must share a slope. So the slope $\frac{b-6}{3-(-2)}$ of the line segment AB should equal the slope $\frac{6}{-2}$ of the line. Therefore, $\frac{b-6}{5} = -3$, which implies that $b = \boxed{-9}$. Final solution: The final answer is -9.

I.4.2 GCSE Exam 2019 — Mathematics Higher

General Certificate of Secondary Education (GCSE) Mathematics is an exam taken annually by approximately 700K students in the UK, with about 40% of students taking the higher tier that we analyze in this evaluation. Students usually take the exam in Year 11, at 16 years old (see the governmental web-page for more details).

For the purpose of this evaluation we focused on the June 2019 exam and identified exam tasks not involving images in an essential way (problems 2, 3, 4, 5, 6b, 7, 8, 10b, 11, 12, 15a, 16, 17, 18, 19, 21, 23, 24, 26, 27). The sum of points possible for all listed problems is 46 out of 80 points for all problems. In our evaluation Minerva 540B obtained 38 out of 46 points (82.6%). In the evaluation we took into account only the top answer in the majority voting.

GSCE Exam 2019 - Selected Samples More straightforward tasks in the GSCE exam (problems that contribute 1 point to the total score) are solved correctly in all instances except one. Some complex problems are also solved correctly; below, we include one straightforward problem and one complex problem together with the model's outputs. We also enclose a sample showing the model can fill a LaTeX table. The last two examples show failure modes: the top solution contains a conceptual error, and the 2nd best choice in majority voting has a calculation mistake, though of a kind that a model can recover from.

Question: Decide which fraction is equivalent to 4.625: $\frac{39}{8}$, $\frac{37}{8}$, $\frac{185}{4}$, or $\frac{17}{4}$.

Model output - 540B: The value of 4.625 is $4.625 = 4 + \frac{625}{1000} = 4 + \frac{5}{8} = \frac{32}{8} + \frac{5}{8} = \frac{37}{8}$. So the answer is $\boxed{\frac{37}{8}}$. Final solution: The final answer is $\frac{37}{8}$.

Question: Solve the simultaneous equations

$$\begin{aligned} 2x + 3y &= 5p \\ y &= 2x + p \end{aligned}$$

where p is a constant.

Model output - 540B: From the second equation, we have $2x = y - p$. Substituting this into the first equation, we get

$$2x + 3y = 5p \text{ i.e., } y - p + 3y \quad = 5p \text{ i.e., } 4y = 6p \text{ i.e., } y \quad = \frac{6p}{4} = \frac{3p}{2}.$$

Substituting $y = \frac{3p}{2}$ into the second equation, we have

$$2x + p = \frac{3p}{2} \text{ i.e., } 2x \quad = \frac{3p}{2} - p \text{ i.e., } 2x = \frac{3p - 2p}{2} \text{ i.e., } 2x \quad = \frac{p}{2} \text{ i.e., } x = \frac{p}{4}.$$

Therefore, the solutions are $x = \frac{p}{4}$ and $y = \frac{3p}{2}$.

Question: Here is some information about the test marks of 120 students.

Mark, m	$0 < m \leqslant 10$	$10 < m \leqslant 20$	$20 < m \leqslant 30$	$30 < m \leqslant 40$	$40 < m \leqslant 50$
Frequency	20	28	40	20	12

Complete the cumulative frequency table. (1 mark)

Mark, m	$m \leqslant 10$	$m \leqslant 20$	$m \leqslant 30$	$m \leqslant 40$	$m \leqslant 50$
Cumulative frequency	20	48			

Model output - 540B: Since 20 students got a mark $0 < m \leq 10$ and 28 students got a mark $10 < m \leq 20$, the cumulative frequency up to $m = 20$ is $20 + 28 = 48$. We are given that 40 students got a mark $20 < m \leq 30$. So the cumulative frequency up to $m = 30$ is $48 + 40 = 88$. We are also given that 20 students got a mark $30 < m \leq 40$. So the cumulative frequency up to $m = 40$ is $88 + 20 = 108$. Finally, 12 students got a mark $40 < m \leq 50$. So the cumulative frequency up to $m = 50$ is $108 + 12 = 120$. We complete the table as follows.

Mark, m	$m \leqslant 10$	$m \leqslant 20$	$m \leqslant 30$	$m \leqslant 40$	$m \leqslant 50$
Cumulative frequency	20	48	88	108	120

Question: Ed and Fay shared £330 in the ratio 7 : 4. Ed gives Fay some of his money. Fay now has the same amount as Ed. How much does Ed give Fay?

Model output - 540B - top choice: First, let us find out how much Ed has initially:

$$\begin{aligned}\text{Ed's money} &= \frac{7}{11} \cdot 330 \\ &= 210,\end{aligned}$$

so

$$\begin{aligned}\text{Fay's money} &= \frac{4}{11} \cdot 330 \\ &= 120.\end{aligned}$$

To find out how much Ed gives Fay, we simply subtract the amount of money that Ed has from the amount of money that Fay has:

$$\begin{aligned}\text{Money given} &= \text{Ed's money} - \text{Fay's money} \\ &= 210 - 120 \\ &= \boxed{90}.\end{aligned}$$

Final solution: The final answer is 90.

Model output - 540B - 2nd best choice: Ed originally has $7x$ and Fay $4x$ for some x. Fay has an extra y after Ed gives her some of his. Then

$$4x + y = 7x - y.$$

We have

$$y = 1.5x.$$

Since Ed has originally

$$330 - 4x = 7x - 330$$

then

$$x = \frac{330}{11} = 30.$$

So Ed gives Fay

$$1.5x = \boxed{45}.$$

Final solution: The final answer is 45.

I.4.3 JEE Main Exam 2020

In our assessment, problems from the JEE exceed the current abilities of the model. Below we feature selected failure cases and leave a complete study for the future.

Question: Compute the limit

$$\lim_{x\to 2} \frac{3^x + 3^{3-x} - 12}{3^{-\frac{x}{2}} - 3^{1-x}}.$$

Model output: We have

$$\begin{aligned}
\lim_{x\to 2} \frac{3^x + 3^{3-x} - 12}{3^{-\frac{x}{2}} - 3^{1-x}} &= \lim_{x\to 2} \frac{3^x + 3^{3-x} - 12}{3^{-\frac{x}{2}} - 3^{1-x}} \cdot \frac{3^{\frac{x}{2}}}{3^{\frac{x}{2}}} \\
&= \lim_{x\to 2} \frac{3^{\frac{3x}{2}} + 3^{\frac{3}{2}} - 12 \cdot 3^{\frac{x}{2}}}{1 - 3^{\frac{x}{2}}} \\
&= \lim_{x\to 2} \frac{3^{\frac{3x}{2}} + 3^{\frac{3}{2}} - 12 \cdot 3^{\frac{x}{2}}}{1 - 3^{\frac{x}{2}}} \cdot \frac{3^{\frac{x}{2}-1}}{3^{\frac{x}{2}-1}} \\
&= \lim_{x\to 2} \frac{3^x + 3^{3-x} - 12}{1 - 3^{1-x}} \\
&= \frac{3^2 + 3^{3-2} - 12}{1 - 3^{1-2}} \\
&= \frac{9 + 3 - 12}{1 - 3} \\
&= \frac{0}{-2} \\
&= \boxed{0}.
\end{aligned}$$

Final solution: The final answer is 0.

Question: If the sum of the coefficients of all even powers of x in the product

$$\left(1 + x + x^2 + x^3 \ldots . + x^{2n}\right)\left(1 - x + x^2 - x^3 \ldots . + x^{2n}\right)$$

is 61. Compute n.

Model output: The coefficient of x^{2k} in the product is $\binom{2n}{2k} - \binom{2n}{2k-2}$. The sum of these coefficients is

$$S = \sum_{k=0}^{n} \binom{2n}{2k} - \binom{2n}{2k-2} = \sum_{k=0}^{n} \binom{2n}{2k} - \sum_{k=0}^{n} \binom{2n}{2k-2} = \sum_{k=0}^{n} \binom{2n}{2k} - \sum_{k=-1}^{n-1} =$$

$$= \binom{2n}{2k} = \binom{2n}{2n} = \boxed{2n+1}.$$

Since $S = 61$, we have $2n + 1 = 61$ and so $n = \boxed{30}$. Final solution: The final answer is 30.

J Memorization Analysis Details

J.1 Dataset overlap

To analyze the degree of overlap between our training data and the MATH evaluation set, we computed the BLEU score of MATH questions and target answers with our training dataset partitioned into 500 word chunks. This overlap computation was performed for the 100 questions and target answers for which the model answered the most confidently (as measured by majority vote fraction). Figure 9 shows the distribution of BLEU scores for these 100 questions.

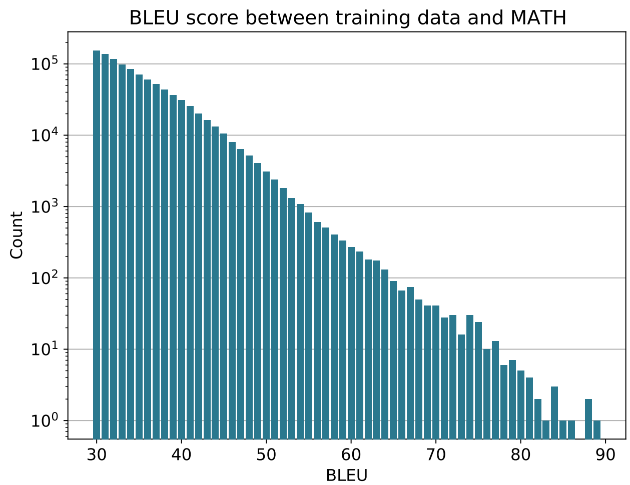

Figure 9: **MATH overlap with training set.** A histogram showing the distribution of BLEU scores between the MATH questions and solutions and the training dataset, cut into 500-word chunks. Samples with BLEU score less than 30 were dropped.

For the 500 most overlapping text segments, we manually inspected the degree of similarity, finding no evidence of dataset contamination. We provide the 500 documents containing these text chunks in the supplementary data.

J.2 Question modifications

To probe the sensitivity of our model to exact problem phrasing we sampled twenty questions that the model answered correctly under majority voting and considered a few varieties of modifications to these questions.

We considered modifications of four types: i) minor modifications to framing, intended to probe weather the model accuracy was solely due to memorizing the exact question statement, ii) modification of the numbers used in problems, iii) larger changes of framing – investigating the models sensitivity to distribution shift, iv) and combinations of number and large framing deformations. In each case, we compared the accuracy of 64 solution samples before and after the modification. The results are shown in Figure 10.

In the case of the two more significant modifications, we see somewhat degraded performance. We note that in those modifications it was more difficult to control for the overall difficulty of the task. We therefore do not interpret this effect as obvious evidence of memorization, but instead present it here to encourage further research.

An example question modification is shown below, and all deformations are provided in the supplementary data.

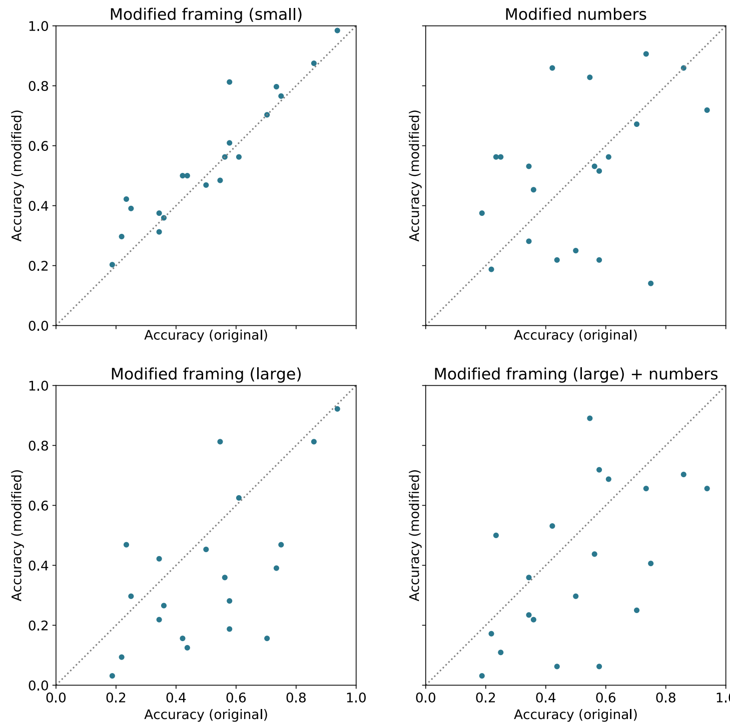

Figure 10: **Performance on modified MATH questions.** Models are evaluated on a 20 question subset of the MATH dataset. Questions are modified by either small framing changes (top-left), number changes (top-right), larger framing changes (bottom-left) or a combination of large framing and number changes (bottom right). Average performance averaged over 64 model solution attempts is plotted on the modified question versus the original question.

Original question: At each basketball practice last week, Jenny made twice as many free throws as she made at the previous practice. At her fifth practice she made 48 free throws. How many free throws did she make at the first practice?

Minor framing modification: At each basketball practice last week, Jenny made twice as many three-pointers as she made at the previous practice. At her fifth practice she made 48 three-pointers. How many three-pointers did she make at the first practice?

Larger framing modification: Roger practiced tennis 10 times last week. In each practice, he served twice as many aces as in the previous practice. During his fifth practice session, he served 48 aces. How many aces did Roger serve in his first practice?

Number modification: At each basketball practice last week, Jenny made thrice as many free throws as she made at the previous practice. At her third practice she made 54 free throws. How many free throws did she make at the first practice?

Combined deformation: Roger practiced tennis 10 times last week. In each practice, he made thrice as many aces as in the previous practice. During his third practice session, he served 54 aces. How many aces did Roger serve in his first tennis practice?

J.3 Solution overlap

As an additional probe of memorization, we investigated the similarity between model generated solutions and ground truth solutions in the MATH dataset. In Figure 11 we show histograms for both the raw BLEU and ROUGE scores (left) and the fraction of examples with BLEU score less than or

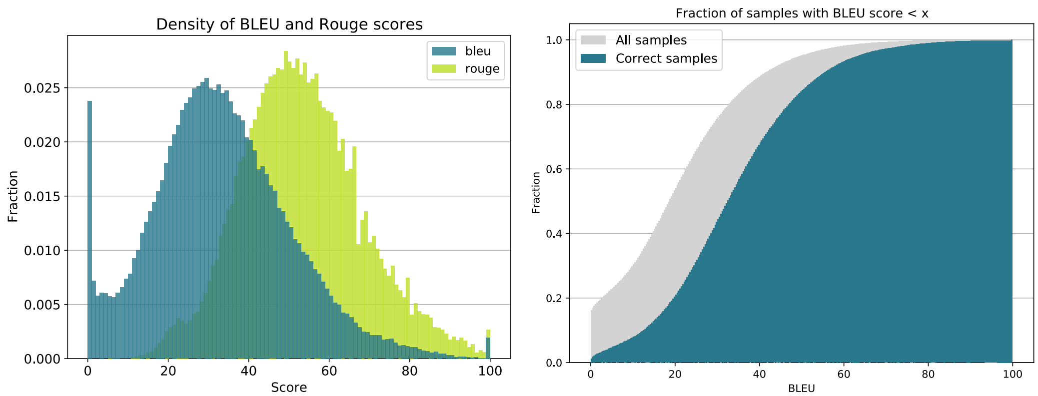

Figure 11: **BLEU and ROUGE scores between model samples and targets**: **(left)** A histogram of the BLEU and ROUGE scores for all correct samples. **(right)** Fraction of samples below a given threshold BLEU score.

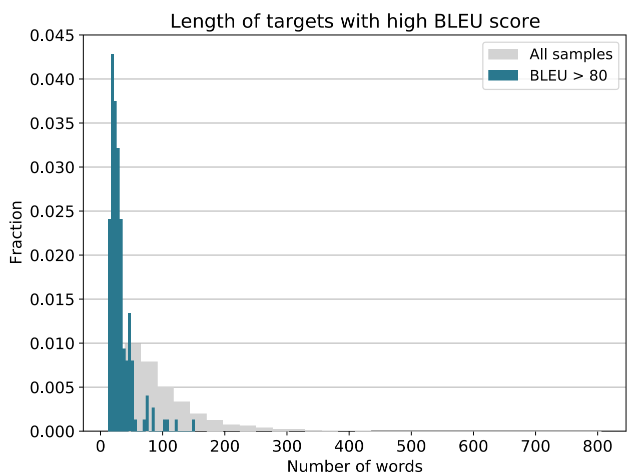

Figure 12: **Samples with high similarity tend to be short**: We show the length distribution (number of words) for samples with high BLEU score (> 80) in blue as compared to all samples in gray. We see that the highly similar samples are skewed towards fewer words.

equal to the x value. Of the 5,000 problems in the test set, we find only 160 problems which have at least one (of 256) samples with BLEU score greater than 80.

We present a few of these examples below and provide all high overlap examples as supplementary data. We note that the examples with high BLEU score tend to be shorter than typical solutions (see Figure 12).

BLEU = 100

Target: We have that $\left|\frac{a}{b}\right| = \frac{|a|}{|b|} = \frac{6}{4} = \boxed{\frac{3}{2}}$.

Model output: We have that $\left|\frac{a}{b}\right| = \frac{|a|}{|b|} = \frac{6}{4} = \boxed{\frac{3}{2}}$.

BLEU = 89.97

Target: We can write

$$\begin{aligned}\frac{\sin x}{\cos x}+\frac{\cos x}{1+\sin x}&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{(1+\sin x)(1-\sin x)}\\&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{1-\sin^2 x}\\&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{\cos^2 x}\\&=\frac{\sin x}{\cos x}+\frac{1-\sin x}{\cos x}\\&=\frac{1}{\cos x}=\boxed{\sec x}.\end{aligned}$$

Model output:

$$\begin{aligned}\frac{\sin x}{\cos x}+\frac{\cos x}{1+\sin x}&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{(1+\sin x)(1-\sin x)}\\&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{1-\sin^2 x}\\&=\frac{\sin x}{\cos x}+\frac{\cos x(1-\sin x)}{\cos^2 x}\\&=\frac{\sin x}{\cos x}+\frac{1-\sin x}{\cos x}\\&=\frac{\sin x+1-\sin x}{\cos x}\\&=\frac{1}{\cos x}\\&=\boxed{\sec x}.\end{aligned}$$

BLEU = 80.02

Target: We have:

$\frac{3}{\sqrt{27}}=\frac{3\sqrt{3}}{\sqrt{81}}=\frac{3\sqrt{3}}{9}=\boxed{\frac{\sqrt{3}}{3}}.$

Model output: We have that

$$\frac{3}{\sqrt{27}}=\frac{3}{3\sqrt{3}}=\frac{1}{\sqrt{3}}=\boxed{\frac{\sqrt{3}}{3}}.$$