

TUNING CHATGPT MATHEMATICAL REASONING LIMITATIONS AND FAILURES WITH PROCESS SUPERVISION

Indrasen Poola

DOI: <https://doi.org/10.5281/zenodo.8296440>

Published Date: 29-August-2023

Abstract: In ChatGPT, vast amounts of text-based data, generally scraped from the public internet, are used to train AI algorithms known as Large Language Models (LLMs). However, LLMs have some severe weaknesses, including the ability to make mistakes, "hallucinate" false information, and bias. They aren't magic, they're not artificial general intelligence, and they have these flaws. The tendency of the models to produce text that seems to be accurate but is actually untrue or not based on the input given is known as hallucination. For instance, even when a language model completely constructed the answer to a question concerning a historical occurrence that never took place, it may nonetheless produce a believable response. Hallucinations are the name for these fabricated LLM responses.

Keywords: ChatGPT, public internet, Large Language Models (LLMs).

INTRODUCTION

ChatGPT doesn't seem to be very good in coming up with original answers to new problems, especially unresolved problems in mathematics. ChatGPT is not a calculator or a math prodigy; it is an AI text-based language model. Large language models now do complex multi-step reasoning far better than they did in the past. But even cutting-edge models can make logical errors, sometimes known as hallucinations. The prevention of hallucinations is a crucial step in the development of aligned AGI. By utilizing either outcome supervision or incentive models, we can train them to recognize hallucinations. Hallucinations are the name for these fabricated LLM responses. ChatGPT doesn't seem to be very good in coming up with original answers to new problems, especially unresolved problems in mathematics.

ChatGPT is not a calculator or a math prodigy; it is an AI text-based language model. Large language models now do complex multi-step reasoning far better than they did in the past. But even cutting-edge models can make logical errors, sometimes known as hallucinations. The prevention of hallucinations is a crucial step in the development of aligned AGI. By utilizing either outcome supervision or incentive models, we can train them to recognize hallucinations. which offers feedback based on the outcome, or process supervision, which offers feedback for each action taken in a series of steps. Using the MATH dataset as our testbed, we thoroughly compare these two approaches based on prior research. Even when success is measured by results, we show that process oversight considerably improves outcomes. We make available our complete process supervision dataset in order to promote related research.

1. METHODS

We compare the process and outcome control of training reward models. While process-supervised reward models (PRMs) receive feedback for each step in the chain of reasoning, outcome-supervised reward models (ORMs) are trained using only the conclusion of the model's chain of reasoning. Process oversight should be favored for several good reasons. Since it identifies the precise position of any faults that take place, it offers more accurate feedback. Additionally, it offers numerous

benefits that are pertinent to AI alignment: it is simpler for people to understand, and it more explicitly rewards models for adhering to a human-endorsed line of reasoning.

2. OUTCOME-SUPERVISED REWARD MODELS (ORMS)

We train ORMs using a strategy similar to Cobbe et al (2021). We randomly select a certain number of answers from the generator for each problem, and we train the ORM to determine whether or not each answer is right or wrong. The final response is typically checked automatically to confirm correctness in practice, but in theory these labels might be supplied by people. The final token prediction made by the ORM is used as the solution's total score at test time. We draw attention to the fact that the automated grading system used to establish ORM objectives is not entirely accurate: false positive solutions that arrive at the right answer through erroneous reasoning will receive a low score.

3. PROCESS-SUPERVISED REWARD MODELS (PRMS)

After the final token in each step, we train PRMs to predict whether the subsequent steps will be right. We maximize the log-likelihood of these target tokens during training in order to represent this prediction as a single token. Therefore, it is sufficient to run a single PRM forward pass across the entire solution to determine the step-level predictions at test time. We display high-resolution PRM scores for two distinct solutions. It is important to calculate a single score for each solution in order to compare numerous answers. This is a crucial but simple point: we define the PRM score for a solution as the likelihood that each step is accurate under the PRM. This is accomplished by combining the odds of each step's correctness. We discuss more score options and further PRM training. When we offer process supervision, we consciously decide to simply supervise the initial mistaken step. Because of this, it is easier to compare outcome and process supervision. Both approaches offer the same information for correct answers, namely that each step was done correctly. Both approaches identify at least one mistake for wrong solutions, and process supervision also identifies the precise place of the error. Process monitoring would have a higher information advantage if we provided extra process supervision after the initial error. This choice also maintains the labelling cost for humans at a similar level: without relying on a final solution that is simple to verify, assessing the correctness of a solution is identical to locating its first error. While the majority of math problems do have straightforward ultimate solutions, we anticipate that this won't hold true in more complicated areas. rained without any extra adjustments into a regular language model pipeline.

4. PROCESS VS OUTCOME SUPERVISION

These three series of reward models are all trained on the same datasets and only change in the choice of supervision. For additional information on the usage of PRM large in outcome and process supervision. We rank each reward model out of 500 and then evaluate it. At all data collecting scales, we observe that process supervision performs much better than both types of outcome supervision. The best reward model from each series is assessed based on its best-of-N performance for various N values. We find that utilizing PRM large for outcome supervision significantly outperforms final-answer checking in terms of efficiency. This can be explained by the fact that PRM large offers superior supervision for solutions that use flawed reasoning to arrive at the right result in the end. Which type of outcome supervision baseline—PRM large or final-answer checking—represents the more suitable outcome monitoring? Although final answer supervision is more explicitly outcome-based, the MATH dataset may have overemphasized its major flaw, the occurrence of false positives. In areas where false positives are less likely to occur, outcome monitoring by PRM large captures outcome supervision better. Although we believe that outcome supervision by PRM as a whole is the most pertinent baseline, we strongly advise readers to come to their own conclusions. Uesato et al. (2022) compare the effects of result and process monitoring in the area of grade school math in a study that is closely analogous to our own. They discovered that the final-answer error rates produced by the two approaches were comparable, and that process supervision produced those outcomes with less data. Our basic concept is fairly similar, however there are three key aspects that are different. First, we execute our large-scale tests using a model that is more capable of collecting the PRM800K dataset. Our small-scale findings in the final section, however, imply that large-scale models are not required to see the advantages of process oversight. The MATH dataset, which is substantially more difficult than GSM8K, is the second one we evaluate on. Third, we gather significantly more information about process supervision. The findings from Uesato et al. (2022) would initially appear to contradict our assertion that process oversight improves performance. However, we think that the disparity in supervisory magnitude can account for the apparent disagreement. A modest quantity of process supervision and a big amount of outcome supervision do, in fact,

result in comparable performance, according to the data scaling trend in, which is in line with Uesato et al (2022). The tendency also demonstrates that, even when evaluated only on results, process monitoring outperforms outcome supervision when scaled up. This agrees with what we found. These findings, in our opinion, strongly support the use of process supervision.

5. SYNTHETIC SUPERVISION

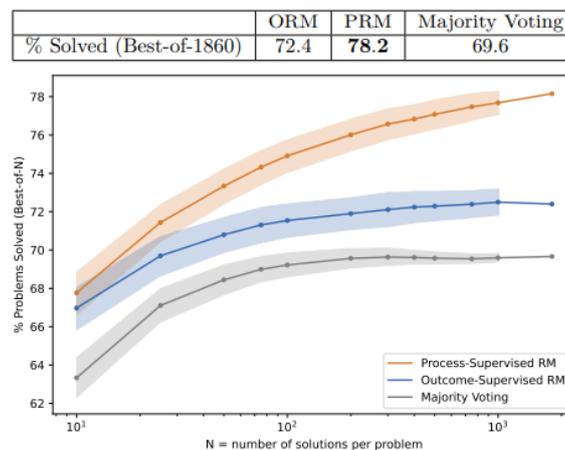
Gao et al. (2022) employ a big reward model to oversee the training of smaller models, which is similar to our approach in Section 4. With tests that call for a significant amount of information about human preferences, they investigate the over-optimization that takes place during RLHF. They utilize a gold-standard reward mechanism in place of human feedback to get around this problem. Similar to their strategy, we employ a large-scale reward model to oversee smaller reward models.

6. ALIGNMENT IMPACT

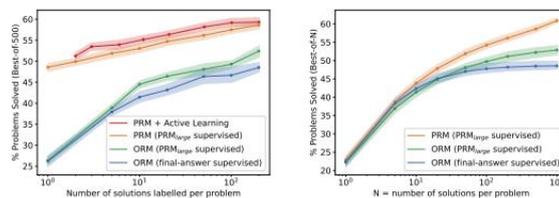
Compared to result supervision, process supervision offers a number of alignment benefits. Since each stage of the process is carefully supervised, the model is explicitly rewarded for adopting a coherent line of reasoning. Since it pushes the model to adhere to a human-approved method, process supervision also increases the likelihood that the reasoning produced will be understandable. In contrast, result supervision is often more difficult to audit and may reward an unaligned process. Safer practices for AI systems might occasionally result in decreased performance, a price known as an alignment tax. Due to pressure to use the most capable model, any alignment tax may generally prevent the adoption of alignment approaches. Our findings below demonstrate that, at least in the arithmetic area, process supervision really leads in a negative alignment tax. This might lead to a greater uptake of process supervision, which would, in our opinion, have favorable alignment side-effects.

7. TECHNICAL DETAILS OF ACTIVE LEARNING

We then look into the effects of active learning. We score 1000 samples from each problem using the small-scale reward model PRM selector, which we trained on one sample from each problem. We choose N samples each problem, of which 80% are the most convincing wrong-answer samples (as determined by the PRM selector), and 20% are the most convincing samples that remain, in order to train each of our larger reward models (right- or wrong-answer). We use PRM large to score the chosen samples, then we train using those scores. This procedure makes sure that every sample is fairly persuasive under the PRM selector, that a significant portion is known to include at least one error, and that the entire dataset is not unduly biased toward incorrect responses. Figure 4a displays how well this data labelling strategy performed. We estimate that this type of active learning is roughly 2.6 times more data efficient than uniform data labelling by comparing the slopes of the line of best fit with and without active learning. We observe that the model appears to slightly underperform the predicted trend line when tested on the largest active learning dataset (200 samples per problem). This discovery is best explained by the fact that 200 samples make up a sizable portion of the whole selection pool (1000 samples), and that this



A comparison of reward models that are process-supervised and outcome-supervised based on how well they can look through a large number of test answers. A solid baseline is demonstrated by majority vote. For N 1000, we display the variance over several subsamples of the 1860 total solutions we produced for each problem..



(a) Four series of reward models trained using different data collection strategies, compared across training sets of varying sizes. (b) Three reward models trained on 200 samples/problem using different forms of supervision, compared across many test-time compute budgets.

	ORM	PRM	Majority Vote	# Problems
AP Calculus	68.9%	86.7%	80.0%	45
AP Chemistry	68.9%	80.0%	71.7%	60
AP Physics	77.8%	86.7%	82.2%	45
AMC10/12	49.1%	53.2%	32.8%	84
Aggregate	63.8%	72.9%	61.3%	234

Recent STEM assessments are used to measure out-of-distribution generalization. Using 100 test samples for each problem, we assess the outcome-supervised RM, the process-supervised RM, and majority voting. We also conducted a preliminary analysis of the effects of repeatedly retraining the PRM selector during data collection. We updated the PRM selector between iterations using all the labelled data available. Unfortunately, this method showed instability that we were unable to identify. The reward models that were produced didn't perform any better than the ones mentioned above. We expect some type of iterative retraining to be helpful in active learning, although there isn't any proof of this just yet. We believe that this is an intriguing area for further study.

8. OUR MAIN CONTRIBUTIONS ARE AS FOLLOWS

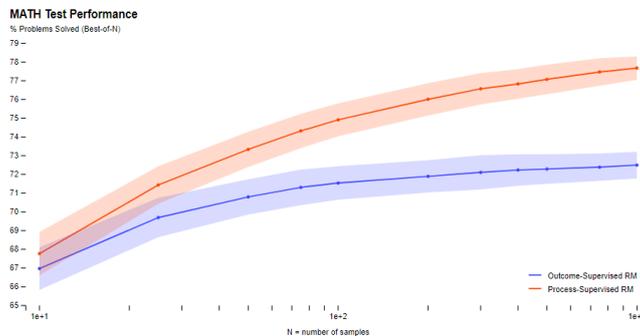
1. We demonstrate that process supervision, as opposed to outcome supervision, can train substantially more trustworthy reward models. We solve 78.2% of the problems from a representative sample of the MATH exam set using our cutting-edge PRM.
2. We demonstrate that a big reward model can effectively carry out large-scale data collection ablations and can reliably imitate human supervision for smaller reward models.
3. We demonstrate that active learning increases process supervision's data efficiency by 2.6 percent.
4. To encourage related study, we make our complete process supervision dataset, PRM800K, available.

Using issues from the MATH test set, we assessed our process-supervised and outcome-supervised reward models. For each issue, we produce a number of potential solutions before selecting the one that each reward model rates as the best. The graph displays, in relation to the number of solutions investigated, the proportion of solutions that are ultimately picked that lead to the right answer. The performance discrepancy worsens as we evaluate more solutions per problem, and the process-supervised reward model not only performs better overall. This demonstrates how much more trustworthy the process-supervised reward model.

We showcase 10 problems and solutions below, along with commentary about the reward model's strengths and weaknesses.

I have found three potential constraints in the chatbot's responses as I work to maximize its potential as an AI-powered chatbot to advance our study. These include (i) limitations on computational calculation, (ii) the possibility of factual inaccuracies, and (iii) insufficient inferential competence. To ensure that the chatbot's response are evaluated and used with the necessary caution and context, it is imperative to be aware of these limits.

Solving MATH problems



Explore examples in 3 categories:

1. True positives

8.1 Problem

Simplify $\tan 100^\circ + 4 \sin 100^\circ \cdot \tan 100^\circ + 4 \sin 100^\circ$
 ChatGPT Model attempt

This difficult trigonometry issue calls for the application of numerous identities in a rather illogical order. Due to the difficulty in determining which identities are truly useful, the majority of solutions are unsuccessful. Although only 1% of solution attempts result in the right answer, the reward model correctly acknowledges that this solution is valid despite the fact that GPT-4 typically cannot solve this problem.

```

1 I notice that the angle 100 degrees is not a special angle, so I will need to use some
2 trigonometric identities to simplify this expression.
3 I recall that  $\tan \theta = \frac{\sin \theta}{\cos \theta}$  and  $\sin 2\theta = 2 \sin \theta \cos \theta$ .
4 I can use these identities to rewrite the expression as  $\frac{\sin 100^\circ}{\cos 100^\circ} + 4 \sin 100^\circ$ .
5 I can combine the terms by finding a common denominator of  $\cos 100^\circ$ .
6 I get  $\frac{\sin 100^\circ + 4 \sin 100^\circ \cos 100^\circ}{\cos 100^\circ}$ .
7 I can simplify the numerator by using the double angle identity again, but in reverse.
8 I get  $\frac{\sin 100^\circ + 2 \sin 200^\circ}{\cos 100^\circ}$ .
9 I notice that 200 degrees is the same as 180 degrees plus 20 degrees, so I can use the sum
10 angle identity for sine.
11 I get  $\frac{\sin 100^\circ + 2(\sin 180^\circ \cos 20^\circ + \cos 180^\circ \sin 20^\circ)}{\cos 100^\circ}$ .
12 I recall that  $\sin 180^\circ = 0$  and  $\cos 180^\circ = -1$ , so I can simplify further.
13 I get  $\frac{\sin 100^\circ - 2 \sin 20^\circ}{\cos 100^\circ}$ .
14 I notice that 100 degrees is the same as 90 degrees plus 10 degrees, so I can use the sum
15 angle identity for sine and cosine.
16 I get  $\frac{(\sin 90^\circ \cos 10^\circ + \cos 90^\circ \sin 10^\circ) - 2 \sin 20^\circ}{(\cos 90^\circ \cos 10^\circ - \sin 90^\circ \sin 10^\circ)}$ .
17 I recall that  $\sin 90^\circ = 1$  and  $\cos 90^\circ = 0$ , so I can simplify further.
18 I get  $\frac{\cos 10^\circ - 2 \sin 20^\circ}{\cos 10^\circ}$ .
19 I can simplify the sign by multiplying the numerator and denominator by -1.
20 I get  $\frac{2 \sin 20^\circ - \cos 10^\circ}{\sin 10^\circ}$ .
21 I notice that 20 degrees is the same as 30 degrees minus 10 degrees, so I can use the
22 difference angle identity for sine.
23 I get  $\frac{2(\sin 30^\circ \cos 10^\circ - \cos 30^\circ \sin 10^\circ) - \cos 10^\circ}{\sin 10^\circ}$ .
24 I recall that  $\sin 30^\circ = \frac{1}{2}$  and  $\cos 30^\circ = \frac{\sqrt{3}}{2}$ , so I can simplify further.
25 I get  $\frac{\cos 10^\circ - \sqrt{3} \sin 10^\circ - \cos 10^\circ}{\sin 10^\circ}$ .
26 I can cancel out the  $\cos 10^\circ$  terms in the numerator.
27 I get  $\frac{-\sqrt{3} \sin 10^\circ}{\sin 10^\circ}$ .
28 I can cancel out the  $\sin 10^\circ$  terms in the numerator and denominator.
29 I get  $-\sqrt{3}$ .
30 Answer:  $-\sqrt{3}$ 
    
```

8.2 Problem

Let $x^8 + 3x^4 - 4 = p_1(x)p_2(x) \cdots p_k(x)$,
 where each non-constant polynomial $p_i(x)$ is monic with integer coefficients,
 and cannot be factored further over the integers. Compute $p_1(1) + p_2(1) + \cdots + p_k(1)$.

ChatGPT Model attempt

Here, GPT-4 completes a challenging series of polynomial factorization with success. The use of the Sophie-Germain identity in step 5 is a crucial and potentially enlightening step.

```

1 I notice that the given polynomial has even degree and only even powers of x, so I can try to
2 make a substitution to simplify it.
3 Let  $y = x^2$ , then the polynomial becomes  $y^2 + 3y - 4$ , which is a quadratic equation.
4 I can factor this quadratic equation as  $(y + 4)(y - 1)$ , so the original polynomial is  $(x^4 +
5 4)(x^2 - 1)$ .
6 Now I need to factor each of these factors further, if possible, over the integers.
7 For the first factor,  $x^4 + 4$ , I recall the Sophie Germain identity, which says that  $a^4 + 4b^4 =
8 (a^2 + 2b^2 + 2ab)(a^2 + 2b^2 - 2ab)$ .
9 Applying this identity with  $a = x$  and  $b = 1$ , I get  $x^4 + 4 = (x^2 + 2x + 2)(x^2 - 2x + 2)$ .
10 For the second factor,  $x^4 - 1$ , I can use the difference of squares formula, which says that
11  $a^2 - b^2 = (a + b)(a - b)$ .
12 Applying this formula with  $a = x^2$  and  $b = 1$ , I get  $x^4 - 1 = (x^2 + 1)(x^2 - 1)$ .
13 Then I can apply the difference of squares formula again with  $a = x$  and  $b = 1$ , to get  $x^2 -
14 1 = (x + 1)(x - 1)$ .
15 So the original polynomial is  $(x^4 + 4)(x^4 - 1) = (x^2 + 2x + 2)(x^2 - 2x + 2)(x^2 +
16 1)(x + 1)(x - 1)$ .
17 These are all the irreducible factors over the integers, so  $p_1(x) = x^2 + 2x + 2$ ,  $p_2(x) =
18 x^2 - 2x + 2$ ,  $p_3(x) = x^2 + 1$ ,  $p_4(x) = x + 1$ , and  $p_5(x) = x - 1$ .
19 To find the sum of their values at  $x = 1$ , I just need to plug in  $x = 1$  into each factor and add
20 them up.
21 I get  $p_1(1) = 1^2 + 2 \cdot 1 + 2 = 5$ ,  $p_2(1) = 1^2 - 2 \cdot 1 + 2 = 1$ ,  $p_3(1) = 1^2 + 1 = 2$ ,
22  $p_4(1) = 1 + 1 = 2$ , and  $p_5(1) = 1 - 1 = 0$ .
23 So the final answer is  $5 + 1 + 2 + 2 + 0 = 10$ .
24 Answer: 10
    
```

8.3 Problem

Find the unique odd integer t such that $0 < t < 23$ and $t + 2$ is the inverse of t modulo 23.

ChatGPT Model attempt

GPT-4 begins conducting guess-and-check in steps 7 and 8. This is a frequent instance where the model may delude itself by asserting that a particular guess was correct when it wasn't. The reward model in this instance validates each step and finds that the chain of reasoning is accurate.

```

1 I need to find an odd integer  $t$  that satisfies the congruence  $t + 2 \equiv t^{-1} \pmod{23}$ .
2 To do that, I can multiply both sides by  $t$  and get  $t^2 + 2t \equiv 1 \pmod{23}$ .
3 Then I can complete the square by adding and subtracting  $(2/2)^2 = 1$  to the left side and get
4  $(t + 1)^2 - 1 \equiv 1 \pmod{23}$ .
5 Simplifying, I get  $(t + 1)^2 \equiv 2 \pmod{23}$ .
6 Now I need to find a square root of 2 modulo 23.
7 I can try some values of  $t + 1$  and see if they work.
8 For example, if  $t + 1 = 3$ , then  $(t + 1)^2 = 9 \not\equiv 2 \pmod{23}$ .
9 If  $t + 1 = 5$ , then  $(t + 1)^2 = 25 \equiv 2 \pmod{23}$ .
10 So one possible value of  $t$  is  $t = 5 - 1 = 4$ .
11 But this is not odd, so I need to find another one.
12 I know that if  $t + 1$  is a square root of 2 modulo 23, then so is  $-t - 1$ .
13 So I can try  $-t - 1 = 5$  and get  $t = -6$ .
14 But this is negative, so I need to add 23 to get a positive value.
15 So another possible value of  $t$  is  $t = -6 + 23 = 17$ .
16 This is odd and satisfies the original congruence, so this is the unique solution.
17 Answer: 17
    
```

8.4 Problem

Find the degree measure of the least positive angle θ for which $\tan \theta = \frac{\cos 5^\circ \cos 20^\circ + \cos 35^\circ \cos 50^\circ - \sin 5^\circ \sin 20^\circ - \sin 35^\circ \sin 50^\circ}{\sin 5^\circ \cos 20^\circ - \sin 35^\circ \cos 50^\circ + \cos 5^\circ \sin 20^\circ - \cos 35^\circ \sin 50^\circ}$

ChatGPT Model Attempt

The expression is successfully made simpler by the model using a number of trigonometric identities.

```

1 I notice that the numerator and denominator of the fraction are both sums of products of
  cosine and sine functions of angles that add up to 25 degrees or 85 degrees.
2 This suggests that I might be able to use some trigonometric identities to simplify the
  expression.
3 For example, I recall that  $\cos(A + B) = \cos A \cos B - \sin A \sin B$  and  $\sin(A + B) =$ 
   $\sin A \cos B + \cos A \sin B$ .
4 If I apply these identities to the terms in the numerator, I get  $\cos 5^\circ \cos 20^\circ +$ 
   $\cos 35^\circ \cos 50^\circ - \sin 5^\circ \sin 20^\circ - \sin 35^\circ \sin 50^\circ = \cos(5^\circ + 20^\circ) + \cos(35^\circ +$ 
   $50^\circ) = \cos 25^\circ + \cos 85^\circ$ .
5 Similarly, if I apply these identities to the terms in the denominator, I get  $\sin 5^\circ \cos 20^\circ -$ 
   $\sin 35^\circ \cos 50^\circ + \cos 5^\circ \sin 20^\circ - \cos 35^\circ \sin 50^\circ = \sin(5^\circ + 20^\circ) - \sin(35^\circ +$ 
   $50^\circ) = \sin 25^\circ - \sin 85^\circ$ .
6 Now I have  $\tan \theta = \frac{\cos 25^\circ + \cos 85^\circ}{\sin 25^\circ - \sin 85^\circ}$ .
7 I wonder if I can simplify this fraction further.
8 I recall another trigonometric identity that says  $\cos A + \cos B = 2 \cos \frac{A+B}{2} \cos \frac{A-B}{2}$ .
9 If I apply this identity to the numerator, I get  $\cos 25^\circ + \cos 85^\circ =$ 
   $2 \cos \frac{25^\circ + 85^\circ}{2} \cos \frac{25^\circ - 85^\circ}{2} = 2 \cos 55^\circ \cos(-30^\circ)$ .
10 I also recall that  $\cos(-\theta) = \cos \theta$ , so I can write  $\cos(-30^\circ) = \cos 30^\circ$ .
11 Similarly, I recall another trigonometric identity that says  $\sin A - \sin B =$ 
   $2 \cos \frac{A+B}{2} \sin \frac{A-B}{2}$ .
12 If I apply this identity to the denominator, I get  $\sin 25^\circ - \sin 85^\circ =$ 
   $2 \cos \frac{25^\circ + 85^\circ}{2} \sin \frac{25^\circ - 85^\circ}{2} = 2 \cos 55^\circ \sin(-30^\circ)$ .
13 I also recall that  $\sin(-\theta) = -\sin \theta$ , so I can write  $\sin(-30^\circ) = -\sin 30^\circ$ .
14 Now I have  $\tan \theta = \frac{2 \cos 55^\circ \cos 30^\circ}{2 \cos 55^\circ (-\sin 30^\circ)} = -\frac{\cos 30^\circ}{\sin 30^\circ}$ .
15 I know that  $\cos 30^\circ = \frac{\sqrt{3}}{2}$  and  $\sin 30^\circ = \frac{1}{2}$ , so I can simplify this to  $\tan \theta = -\frac{\sqrt{3}}{1} = -\sqrt{3}$ .
16 I also know that  $\tan \theta$  is negative in the second and fourth quadrants, and that  $\tan \theta = -\sqrt{3}$ 
  when  $\theta$  is  $120^\circ$  or  $300^\circ$ .
17 Therefore, the least positive angle  $\theta$  for which  $\tan \theta = -\sqrt{3}$  is  $120^\circ$ .
18 Answer: 120
    
```

9. 2. True Negatives

9.1 Problem

Recently, an exam was taken by a class of 30 students. What was the class mean (mean) on this test if 20 students scored 80, 8 students scored 90, and 2 students got 100?

Attempt at ChatGPTModel

GPT-4 attempts to incorrectly simplify an expression in step seven. This error is detected by the reward model.

```

1 To find the class average, I need to add up all the scores and divide by the number of students.
2 I can use a shortcut to add up the scores, since they are grouped by value.
3 For example, instead of adding  $80 + 80 + \dots + 80$  twenty times, I can just multiply 80 by 20 to
  get the same result.
4 Similarly, I can multiply 90 by 8 and 100 by 2 to get the sums of those groups.
5 So, the total sum of the scores is  $80 * 20 + 90 * 8 + 100 * 2$ .
6 I can use the distributive property of multiplication to simplify this expression.
7 It is equivalent to  $80 * (20 + 1) + 90 * 8 + 100$ .
8 That is,  $80 * 21 + 720 + 100$ .
9 I can do the arithmetic in my head or use a calculator.
10  $80 * 21$  is 1680, so the total sum is  $1680 + 720 + 100$ , which is 2500.
11 Now, to find the average, I need to divide 2500 by 30, the number of students.
12 I can use long division or a calculator for this step.
13 2500 divided by 30 is 83.33, repeating.
14 This means that the class average is 83.33, or 83 and one-third.
15 Answer: 83.33
    
```

9.2 Problem

What is the sum of A's unique prime divisors if A is the product of 500500's positive divisors? An effort to use ChatGPT

GPT-4 commits a straightforward calculation error on step 11. This error is detected by the reward model.

```

1 I notice that  $500 = 2^2 \cdot 5^3$ , so its positive divisors are of the form  $2^a \cdot 5^b$ , where  $a \in \{0, 1, 2\}$  and  $b \in \{0, 1, 2, 3\}$ .
2 To find the sum of these divisors, I can use a formula that involves the geometric series: if  $p$  is a prime number and  $n$  is a nonnegative integer, then the sum of the powers of  $p$  from 0 to  $n$  is  $\frac{p^{n+1}-1}{p-1}$ .
3 So, the sum of the powers of 2 from 0 to 2 is  $\frac{2^{2+1}-1}{2-1} = 7$ , and the sum of the powers of 5 from 0 to 3 is  $\frac{5^{3+1}-1}{5-1} = 156$ .
4 By the distributive property, the sum of the divisors of 500 is the product of these two sums:  $A = 7 \cdot 156 = 1092$ .
5 Now, I need to find the sum of the distinct prime divisors of  $A$ .
6 I can try to factor  $A$  into primes by trial division.
7 I notice that  $A$  is divisible by 2, so I divide by 2 and get 546.
8 I repeat this process and divide by 2 again, getting 273.
9 Now, 273 is not divisible by 2, so I try the next prime, 3.
10 I see that 273 is divisible by 3, so I divide by 3 and get 91.
11 I divide by 3 again, getting 31.
12 Now, 31 is a prime number, so I have found all the prime factors of  $A$ :  $A = 2^2 \cdot 3^2 \cdot 31$ .
13 The sum of the distinct prime divisors of  $A$  is  $2 + 3 + 31 = 36$ .
14 Answer: 36
    
```

9.3 Problem

Let $x^8 + 3x^4 - 4 = p_1(x)p_2(x) \cdots p_k(x)$, where each non-constant polynomial $p_i(x)$ is monic with integer coefficients, and cannot be factored further over the integers. Compute $p_1(1) + p_2(1) + \cdots + p_k(1)$.

ChatGPT Model attempt

In step 12, GPT-4 makes the mistake of attempting to apply the difference of squares formula to an expression that isn't actually a difference of squares. This error is detected by the reward model.

```

1 This problem is about factoring a polynomial over the integers, which can be a tricky task.
2 One possible strategy is to look for rational roots of the polynomial, using the rational root theorem.
3 However, that would involve checking a lot of possible factors of  $-4$ , and there is no guarantee that the polynomial has any rational roots at all.
4 Another possible strategy is to use some algebraic tricks to simplify the polynomial or reduce its degree.
5 I notice that the polynomial has only even powers of  $x$ , so I can try to substitute  $y = x^2$  and see what happens.
6 Then the polynomial becomes  $y^4 + 3y^2 - 4$ , which is a quadratic equation in  $y^2$ .
7 I can try to factor this quadratic equation by looking for two numbers that multiply to  $-4$  and add to 3.
8 Those numbers are 4 and  $-1$ , so I can write  $y^4 + 3y^2 - 4 = (y^2 + 4)(y^2 - 1)$ .
9 Now I can substitute back  $x^2 = y$  and get  $x^8 + 3x^4 - 4 = (x^4 + 4)(x^4 - 1)$ .
10 This looks promising, because I have reduced the degree of the factors from 8 to 4.
11 I can try to factor each of these factors further, using the difference of squares formula.
12 I get  $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x^2 - 1)$ .
13 I can apply the difference of squares formula again to the last factor and get  $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x + 1)(x - 1)$ .
14 Now I have factored the polynomial completely into monic linear and quadratic factors with integer coefficients.
15 These are the  $p_i(x)$ 's that the problem is asking for.
16 To find the sum of their values at  $x = 1$ , I just need to plug in  $x = 1$  into each factor and add them up.
17 I get  $p_1(1) + p_2(1) + \cdots + p_k(1) = (1^2 + 2)(1^2 - 2)(1^2 + 1)(1 + 1)(1 - 1)$ .
18 Simplifying, I get  $p_1(1) + p_2(1) + \cdots + p_k(1) = (3)(-1)(2)(2)(0)$ .
19 Multiplying, I get  $p_1(1) + p_2(1) + \cdots + p_k(1) = 0$ .
20 Answer: 0
    
```

9.4 Problem

Find the unique odd integer t such that $0 < t < 23$ and $t + 2$ is the inverse of t modulo 23.

ChatGPT Model attempt

Step 8's reasoning is odd, but the reward model overlooks it. But in step 9, the model factors the expression erroneously. This error is detected by the reward model.

```

1 I need to find an odd integer  $t$  between 0 and 23 such that  $t + 2$  is the inverse of  $t$  modulo 23.
2 That means that  $t$  and  $t + 2$  multiply to give 1 modulo 23.
3 I can write this as  $t(t + 2) \equiv 1 \pmod{23}$ .
4 I can try to guess and check values of  $t$  until I find one that works, but that might take a long time.
5 I can also try to use some properties of modular arithmetic to simplify the equation.
6 For example, I can subtract 23 from  $t + 2$  and still get the same remainder modulo 23.
7 So I can write  $t(t + 2) \equiv t(t - 21) \pmod{23}$ .
8 This looks simpler, because now I have a difference of two multiples of  $t$ .
9 I can factor out  $t$  and get  $t(t - 21) \equiv t(1 - 21) \pmod{23}$ .
10 Now I can divide both sides by  $t$ , as long as  $t$  is not a multiple of 23.
11 This gives me  $t - 21 \equiv 1 - 21 \pmod{23}$ .
12 I can simplify this further by adding 21 to both sides.
13 This gives me  $t \equiv 1 \pmod{23}$ .
14 This means that  $t$  has the same remainder as 1 when divided by 23.
15 The smallest positive value of  $t$  that satisfies this is  $t = 1$ .
16 This is also an odd integer, so it meets the criteria of the problem.
17 I can check that  $t = 1$  works by plugging it into the original equation.
18 I get  $1(1 + 2) \equiv 1 \pmod{23}$ , which is true.
19 Therefore, the unique odd integer  $t$  that satisfies the given conditions is  $t = 1$ .
20 Answer: 1
    
```

10. 3. False Positives

10.1 Problem

Given that $x_1 = 211, x_2 = 375, x_3 = 420, x_4 = 523$, and $x_n = x_{n-1} - x_{n-2} + x_{n-3} - x_{n-4}$ when $n \geq 5$, find the value of $x_{531} + x_{753} + x_{975}$.

ChatGPT Model attempt

Step 4 of GPT-4 incorrectly states that the sequence repeats every 12 terms when it actually repeats every 10 terms. The reward model can occasionally be deceived by this kind of counting error.

```

1 This is a problem involving a recursive sequence, where each term depends on the previous four terms.
2 One way to approach this problem is to look for a pattern or a formula that can help us find any term of the sequence without having to compute all the previous terms.
3 To do that, let's write down the first few terms of the sequence and see if we notice anything:
 $x_1 = 211,$ 
 $x_2 = 375,$ 
 $x_3 = 420,$ 
 $x_4 = 523,$ 
 $x_5 = x_4 - x_3 + x_2 - x_1 = 523 - 420 + 375 - 211 = 267,$ 
 $x_6 = x_5 - x_4 + x_3 - x_2 = 267 - 523 + 420 - 375 = -211,$ 
 $x_7 = x_6 - x_5 + x_4 - x_3 = -211 - 267 + 523 - 420 = -375,$ 
 $x_8 = x_7 - x_6 + x_5 - x_4 = -375 - (-211) + 267 - 523 = -420,$ 
 $x_9 = x_8 - x_7 + x_6 - x_5 = -420 - (-375) + (-211) - 267 = -523,$ 
 $x_{10} = x_9 - x_8 + x_7 - x_6 = -523 - (-420) + (-375) - (-211) = -267,$ 
 $x_{11} = x_{10} - x_9 + x_8 - x_7 = -267 - (-523) + (-420) - (-375) = 211,$ 
 $x_{12} = x_{11} - x_{10} + x_9 - x_8 = 211 - (-267) + (-523) - (-420) = 375,$ 
 $x_{13} = x_{12} - x_{11} + x_{10} - x_9 = 375 - 211 + (-267) - (-523) = 420,$ 
 $x_{14} = x_{13} - x_{12} + x_{11} - x_{10} = 420 - 375 + 211 - (-267) = 523.$ 
4 We can see that the sequence repeats itself every 12 terms, starting from  $x_1$ .
5 This means that we can find any term of the sequence by finding its remainder when divided by 12, and then looking up the corresponding term in the first cycle.
6 For example, to find  $x_{531}$ , we can divide 531 by 12 and get a quotient of 44 and a remainder of 3.
7 This means that  $x_{531}$  is the same as  $x_3$ , which is 420.
8 Similarly, we can find  $x_{753}$  and  $x_{975}$  by dividing them by 12 and looking up their remainders.
9 We get that  $x_{753}$  has a remainder of 9, so it is the same as  $x_9$ , which is -523.
10 And  $x_{975}$  has a remainder of 3, so it is the same as  $x_3$ , which is 420.
11 Therefore, the value of  $x_{531} + x_{753} + x_{975}$  is  $420 + (-523) + 420 = 317$ .
12 Answer: 317
    
```

10.2 Problem

Denali and Nate work for a dog walking business and are paid for each dog they walk. Denali is responsible for 16 dogs and Nate is responsible for 12 dogs. Under the company's new policy, they will be assigned or unassigned new dogs in groups of x dogs. The ratio of Denali's pay to Nate's pay would be the same if Denali started walking $4x$ more dogs and Nate stayed at 12 dogs or if x of Nate's dogs were reassigned to Denali. Find x if $x \neq 0$.

ChatGPT Model attempt

By grouping like terms in step 13, GPT-4 tries to make the problem simpler. The linear terms are moved and combined correctly to the left, but the right side is accidentally left unaffected. This error deceives the reward model.

```

1 I need to find a value of  $x$  that makes the ratio of Denali's pay to Nate's pay the same in two
2 different scenarios.
3 I can assume that the pay per dog is constant, so the ratio of their pay is the same as the ratio
4 of the number of dogs they walk.
5 Let me write an equation for the first scenario, where Denali walks  $4x$  more dogs and Nate
6 stays at 12 dogs.
7 The ratio of Denali's dogs to Nate's dogs is  $\frac{16+4x}{12}$ .
8 Now let me write an equation for the second scenario, where  $x$  of Nate's dogs are reassigned
9 to Denali.
10 The ratio of Denali's dogs to Nate's dogs is  $\frac{16+x}{12-x}$ .
11 I want these two ratios to be equal, so I can set them equal to each other and solve for  $x$ .
12 I get  $\frac{16+4x}{12} = \frac{16+x}{12-x}$ .
13 I can cross-multiply and simplify this equation.
14 I get  $(16 + 4x)(12 - x) = (16 + x)(12)$ .
15 Expanding the brackets, I get  $192 - 16x + 48x - 4x^2 = 192 + 12x$ .
16 Subtracting 192 from both sides, I get  $-16x + 48x - 4x^2 = 12x$ .
17 Rearranging the terms, I get  $-4x^2 + 20x = 12x$ .
18 Subtracting  $12x$  from both sides, I get  $-4x^2 + 8x = 0$ .
19 I can factor out  $-4x$  from both terms, and get  $-4x(x - 2) = 0$ .
20 This means that either  $-4x = 0$  or  $x - 2 = 0$ .
21 Solving for  $x$ , I get  $x = 0$  or  $x = 2$ .
22 However, the problem says that  $x \neq 0$ , so I can ignore that solution.
23 Therefore, the only valid solution is  $x = 2$ .
24 Answer: 2
    
```

10.3 Problem

When converting $\frac{31}{11111}$ to a decimal, the decimal turns out to be a repeating decimal. How many digits repeat in this repeating decimal? For example, if you get the repeating decimal $0.\overline{123}$, then your answer should be 3, and if you get $0.436\overline{7}$, your answer should be 1.

ChatGPT Model attempt

In step 16, while attempting to do long division, GPT-4 makes the mistake of omitting the leading zeros from the repeating portion of the decimal. This error deceives the reward model.

```

1 To convert a fraction to a decimal, I can divide the numerator by the denominator using long
  division.
2 I'll start by writing 31 over 11111 and bring down a zero.
3 Then I'll ask, how many times does 11111 go into 310?
4 Well, it doesn't go in evenly, so I'll write a zero above the decimal point and bring down another
  zero.
5 Now I have 3100.
6 How many times does 11111 go into 3100?
7 Still zero, so I'll write another zero above the decimal point and bring down another zero.
8 Now I have 31000.
9 How many times does 11111 go into 31000?
10 Ah, now I can see that it goes in twice, so I'll write a 2 above the decimal point and subtract
  22222 from 31000 to get 8778.
11 I'll bring down another zero and ask, how many times does 11111 go into 87780?
12 It goes in seven times, so I'll write a 7 above the decimal point and subtract 77777 from 87780
  to get 10003.
13 I'll bring down another zero and ask, how many times does 11111 go into 100030?
14 It goes in nine times, so I'll write a 9 above the decimal point and subtract 99999 from 100030
  to get 31.
15 Wait a minute, I'm back to where I started! That means the decimal will repeat from here on.
16 So I'll put a bar over the 279 to indicate that it repeats indefinitely.
17 The decimal is 0.00279 with a bar over 279.
18 The question asks how many digits repeat in this repeating decimal.
19 I can see that there are three digits that repeat: 2, 7, and 9.
20 So the answer is 3.
21 Answer: 3
    
```

10.4 Problem

Each of Bob and Alice's bags holds one ball of the hues blue, green, orange, red, and violet. Bob's bag receives one ball from Alice, which she chooses at random. Bob then takes a ball at random from his bag and places it in Alice's bag. What is the likelihood that the contents of the two bags will be identical following this procedure?

ChatGPT Model attempt

In step 9, GPT-4 commits a minute counting error. On the surface, it seems logical to assert that there are five ways to exchange a ball of the same color because there are five different colors. This undercounts by a factor of 2, though, because Bob has an option between two balls to give Alice. This error deceives the reward model.

```

1 To find the probability of an event, I need to divide the number of favorable outcomes by the
  number of possible outcomes.
2 In this case, the event is that the contents of the two bags are the same after the process.
3 So I need to count how many ways this can happen, and how many ways the process can
  happen in total.
4 I will start with the total number of ways the process can happen.
5 Alice can choose any of the five balls in her bag, and Bob can choose any of the six balls in his
  bag.
6 So the total number of ways the process can happen is 5 times 6, which is 30.
7 Now I will count how many ways the contents of the two bags can be the same after the
  process.
8 This means that Alice and Bob must have exchanged the same color of ball.
9 There are five colors, so there are five ways they can exchange the same color of ball.
10 Therefore, the number of favorable outcomes is 5.
11 So the probability of the event is 5 divided by 30, which simplifies to 1/6.
12 Answer: 1/6
    
```

11. CONCLUSION

In the area of mathematical reasoning, we have demonstrated that process supervision may be used to train reward models that are substantially more accurate and resistant to failures of LLMs and hallucinations than outcome supervision. By selectively revealing the most useful model completions for human feedback, we have also demonstrated how active learning may be utilized to reduce the cost of collecting human data. We make PRM800K, the complete dataset of user feedback used to train our cutting-edge reward model, available in the hopes that doing so will spur additional study on the alignment of big language models. Process supervision is, in our opinion, currently understudied, and we look forward to future research that will examine the degree to which these techniques generalize in greater detail. We believe it is crucial for future work to investigate the impact of process supervision in other fields because it is unclear how broadly these results will extend outside of the realm of mathematics. If these findings hold true across the board, we might discover that process supervision offers us the best of both worlds—a strategy that is both more effective and more in line with the objectives than outcome monitoring.

For the purpose of verifying the sample codes, the sample dataset is given to the general public.

<https://github.com/indrasenp/prm800k>

REFERENCES

- [1] Uesato, J., Kushman N., Kumar R., Song F., Siegel N., Wang L., Creswell A., Irving G. and Higgins, I., 2022. Solving math word problems with process- and outcome-based feedback.
- [2] Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E., Song D. and Steinhardt J., 2021. Measuring Mathematical Problem Solving With the MATH Dataset.
- [3] Ouyang L., Wu J., Jiang X., Almedia D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Aspell A., Welinder P., Christiano P., Leike J. and Lowe R., 2022. Training language models to follow instructions with human feedback
- [4] Creswell, M. Shanahan, and I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning.
- [5] Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al. Solving quantitative reasoning problems with language models.
- [6] Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models.
- [7] Stuhlmüller and J. Byun. Supervise process, not outcomes. <https://ought.org/updates/2022-04-06-process>, 2022
- [8] Zelikman, Y. Wu, J. Mu, and N. Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35: 15476–15488, 2022.
- [9] Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021