Twitter Thread by alewkowycz



alewkowycz
@alewkowycz

Very excited to present Minerva■: a language model capable of solving mathematical questions using step-by-step natural language reasoning. Combining scale, data and others dramatically improves performance on the STEM benchmarks MATH and MMLU-STEM. https://t.co/bQJOyMSCD4

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y-coordinate of the point where this line crosses the y-axis?

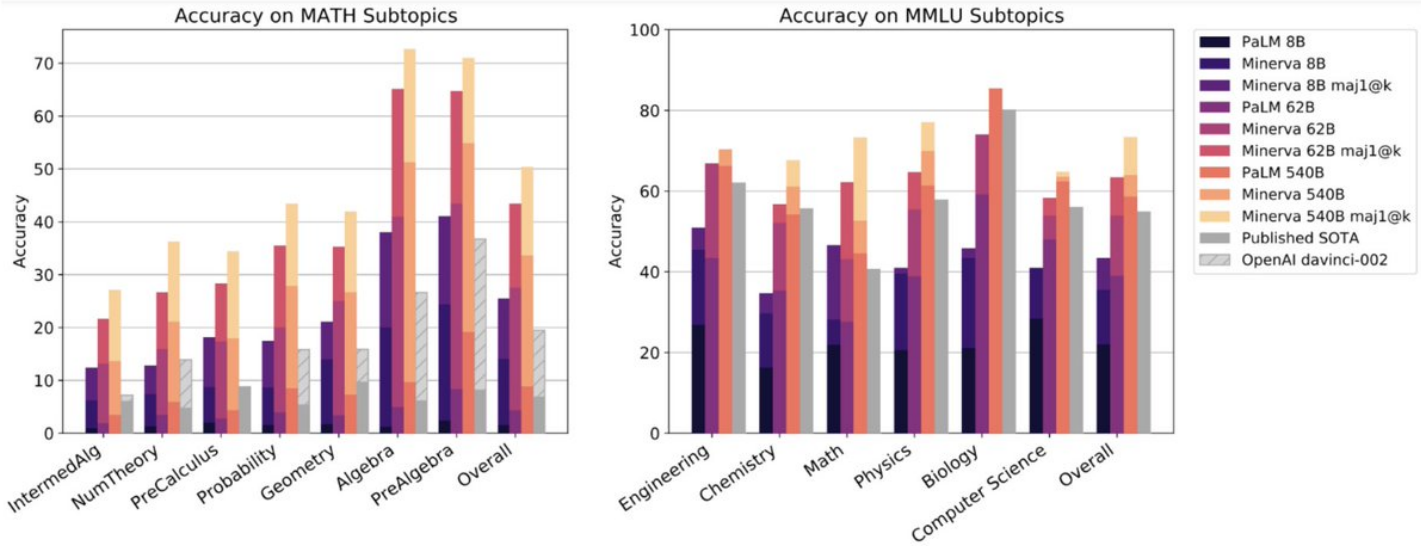
Model output: The line is parallel to $y = 4x+6$, which means that it has the same slope as $y = 4x + 6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$y - 10 = 4(x - 5)$$
$$y - 10 = 4x - 20$$
$$y = 4x - 10$$

Now that we have the equation of the line, we can find the y-coordinate of the point where the line crosses the y-axis by substituting $x = 0$ into the equation:

$$y = 4 \cdot 0 - 10 = \boxed{-10}.$$





Starting from PaLM■, Minerva was trained on a large dataset made of webpages with mathematical content and scientific papers. At inference time, we used chain-of-thought/scratchpad and majority voting to boost performance without the assistance of external tools.





Models' mistakes are easily interpretable. Despite impressive arithmetic skills it still makes calculation mistakes. We estimate the false positive rate due to achieving the right answer from incorrect reasoning and find it relatively low. More samples: https://t.co/YFvsFd6kZ1!

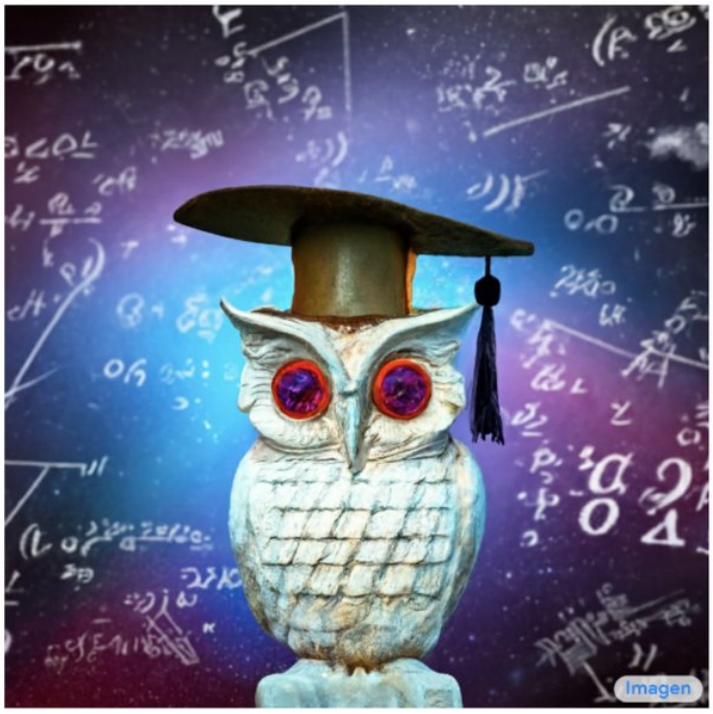
Question: From $1, 2, 3, 4, 5, 6, 7, 8, 9$ we randomly draw with replacement numbers n and m. Compute the probability that $n \times m = 24$.

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 $\frac{4}{81}$



Evaluating our model in 2022 Poland's National Math Exam it performed above the national average, it solved more than 80% GCSE Higher Mathematics problems and evaluated a variety of STEM undergraduate problems from MIT, it solved nearly a third of them.





Paper: https://t.co/W8t2m02UO8

Great collaboration with Anders, @dmdohan, @ethansdyer, @hmichalewski, @vinayramasesh, @AmbroseSlone, @cem__anil, Imanol, Theo, @Yuhu_ai_, @bneyshabur, @guygr and @vedantmisra !

Solving Quantitative Reasoning Problems with Language Models

Aitor Lewkowycz*[1], Anders Andreassen†[1], David Dohan†[2], Ethan Dyer†[1], Henryk Michalewski†[2], Vinay Ramasesh†[1], Ambrose Slone[1], Cem Anil[1], Imanol Schlag[1], Theo Gutman-Solo[3], Yuhuai Wu[3], Behnam Neyshabur*[1], Guy Gur-Ari*[1], and Vedant Misra*[1]

[1]Google Research, Blueshift Team
[2]Google Research, Brain Team
[3]Google Research

Abstract

Language models have achieved remarkable performance on a wide range of tasks that require natural language understanding. Nevertheless, state-of-the-art models have generally struggled with tasks that require quantitative reasoning, such as solving mathematics, science, and engineering problems at the college level. To help close this gap, we introduce Minerva, a large language model pretrained on general natural language data and further trained on technical content. The model achieves state-of-the-art performance on technical benchmarks without the use of external tools. We also evaluate our model on over two hundred undergraduate-level problems in physics, biology, chemistry, economics, and other sciences that require quantitative reasoning, and find that the model can correctly answer nearly a third of them.

#minerva

This was a collaboration between multiple teams at Google Research including Blueshift, Brain, and others. See https://t.co/ttf5HvXSG8 for more info about our team.