International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Is Google's New AI As Smart As A Human?

Barath Maheswaran, Bhuvan Shridhar
BMS College of Engineering, Bengaluru, Karnataka, India

Article Info

Publication Issue :
Volume 10, Issue 1
January-February-2023
Page Number : 144-148

Article History
Accepted : 10 Jan 2023
Published: 27 Jan 2023

ABSTRACT

Natural language understanding tasks have seen impressive results from language models employed in them. 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. In order to bridge the gap, we present Minerva, a big language model trained with standard natural language data and further honed with technical material. The model achieves the best possible results on technical tests without the requirement of any external tools. To assess our model, we have tested it on more than 200 queries from undergraduate-level courses in physics, biology, chemistry, economics, and other sciences that call for quantitative thinking. We have observed that the model is able to accurately respond to nearly one-third of them.

Keywords : Artificial Neural Networks

I. INTRODUCTION

Artificial neural networks have seen remarkable success in a variety of domains including computer vision, speech recognition, audio and image generation, translation, game playing, and robotics. Big language models have demonstrated impressive outcomes on a wide range of natural language activities, including sensible logic, answering questions, and summarizing. Nevertheless, these models have not been successful in tasks that involve quantitative thinking, such as resolving math, science, and engineering issues.

Language models have an intriguing use in quantitative reasoning problems as they test a model's ability in multiple ways. It requires the solver to comprehend the natural language input, recall any relevant details, and perform an algorithm or multiple calculations to get the right answer. Additionally, it is necessary for the solver to completely comprehend and generate precise mathematical symbols and numerals, as well as making use of a computation process to make changes to the symbols or numerals. Finally, this type of problem offers an opportunity to research and develop strong quantitative reasoning solvers that can serve as a helpful resource for humans in scientific and technical fields.

Previous studies have demonstrated that when big language models have been trained on data particular to a certain domain, they have demonstrated exceptional results when applied to mathematical and coding questions.

For this research, we tested this technique on quantitative reasoning issues, wherein the model must give a thorough and independent answer, without using any outside devices. These tasks include mathematics word problems, competition mathematics assessments, and several difficulties connected to science and engineering.

.

1.1 Features

We are introducing Minerva, an advanced language model that has demonstrated impressive results on a variety of quantitative reasoning tasks. This model is capable of interpreting natural language questions that involve scientific and mathematical topics and responding with step-by-step solutions using the correct LATEX notation. As illustrated in the figure, here are examples of Minerva's replies to inquiries concerning mathematics and physics.

Minerva is developed using the PaLM general language models, which are then further refined by training them on a large collection of scientific and mathematical data. To begin, 8B, 62B, and 540B parameter models are used and trained on a technical content dataset. We have seen remarkable results in Math, GSM8k, and a subset of MMLU that include natural language questions related to mathematics and science. It is also worth mentioning that our models work well even when there is only a few training data available, and they don't need to be specifically trained on the evaluation datasets.

This research paper presents an original contribution in the form of a large dataset that combines both natural language and the proper usage of formal mathematical language, including equations and diagrams. The dataset is gathered from the arXiv preprint server and web pages which have been carefully treated to reduce the loss of mathematical information. The results of this work have set a new benchmark for the performance that can be achieved on quantitative reasoning tests by enlarging the quality of data and the size of the model.

To broaden the evaluation of quantitative aptitude, we assembled a collection of over two hundred college-level questions in mathematics and science from MIT's OpenCourseWare(OCW). This provides an assessment of our model's quantitative reasoning skills in a sequence of thought setting beyond a purely mathematical landscape.

II. TRAINING

A) TRAINING DATASET

To create the models, we used a dataset containing 38.5 billion tokens from webpages with mathematical content, as well as papers on the arXiv preprint server. We also included general natural language data, which was the same data used to pre-train PaLM. We constructed our mathematical webpage dataset by collecting pages having mathematical expressions in MathJax format. We cleaned the pages, keeping mathematical notation, such as LATEX symbols and formatting, while removing HTML tags. This allowed the model to process equations like eπi + 1 = 0 and E = mc2 during training.

B) TRAINING PROCEDURE

Our method of tackling this problem is to initially train the PaLM transformer language models without the decoder, and then further develop them with our mathematics dataset using autoregressive objectives. This includes the chief model and training hyper parameters. The biggest model, with 540B parameters, was improved on 26B tokens, even though it is not as trained as the 8B and 62B models, it still presents enhanced performance.

C) EVALUATION OF DATASETS

We mainly focus on few shot evaluation. For evaluation, we truncate the inputs from the left to 1024 tokens and we use the model to generate up to 512 tokens. When sampling once per problem, we sample greedily. When sampling multiple times per problem we use nucleus sampling with temperature T = 0.6, p = 0.95. For generative tasks, the model produces a chain-of-thought answer and demarcates a final answer. We evaluate a solution as correct if the final answer matches the ground truth solution, independent of the quality of the chain-of-thought preceding it. To evaluate correctness, we parse the final answers and compare them using the SymPy library. This is done in order to correctly identify answers that are mathematically equivalent such as $1/\sqrt{3}$ and $\sqrt{3}/3$.

We have implemented this method in our solver for problems such as finding rational approximations for functions like pi or π and solving systems of linear equations like $x + 2y – 5z = 10/2 + 15/2 – 20/2$

We present a model capable of solving 12K middle school and high school mathematics problem statements. The model is trained with a fixed 4-shot prompt, which includes four random examples from the training dataset whose ground truth targets are not too long.

In this work, we present a model that can solve middle school math word problems. Our model is evaluated using the chain-of-thoughts prompt. Previous models evaluated on GSM8k made use of an external calculator. In this work, our model does not have access to any external tools.

MMLU-STEM is a subset of the MMLU dataset focused on science, technology, engineering, and mathematics (STEM). The original version of this dataset was used for training and development tasks. We consider chain-of-thought prompting for this task, where we prompt the model with examples that include step-by-step solutions. We use a multiple-choice version of the MATH prompt for topics that involve mathematical reasoning, and add step-by step solutions to the standard 5-shot prompts for the rest of the topics.

D) STEM-DATASETS

We evaluated the scientific reasoning capabilities of Minerva by collecting a set of problems from OCW courses on solid-state chemistry, differential equations and special relativity. The set of problems was collected using SymPy and a SymPy script that generates OCW Courses files for all MIT OCW courses. Problems were selected based on their difficulty level as measured by the number of steps required to solve them. In order to evaluate the scientific reasoning capabilities of Minerva , we harvested a set of STEM problems at the undergraduate level, most of which involve multi-step reasoning, which we refer to in this paper as OCW Courses . Using publicly-available course materials offered by MIT (OpenCourseWare), we collected problems with automatically-verifiable solutions (either numeric or symbolically verifiable via SymPy) from courses including “solid-state chemistry”, “information and entropy”, “differential equations”, and “special relativity.” These problems were processed by contractors to be self-contained and to have a clearly-delineated final answer. Problems asking for a proof or open-ended short answer were not included. In total we curated 272 problems, 191 of which have numeric solutions and 81 have symbolic solutions.

E) INFERENCE-TIME METHOD

When faced with a problem, we can generally find that there are many different wrong ways to solve it but only one or two right ones. This is why when faced with such problems, choosing the best solution among many incorrect ones becomes important. In contrast to pass@k, where a problem is considered solved if any single answer works out of k attempts, maj1@k consists of splitting answers based on how they turn out and picking the most popular answer among those groups. Logically speaking, this makes sense because while there may be various wrong methods for solving a problem, there will only ever be just one or two correct methods--this also means that as you increase k (the number of tries), you'll need more time before being able to make an accurate decision. In contrast though, because of how pass@k works--where its ability relies on what might happen in the future--it doesn't require nearly as much time or energy before coming up with an accurate decision.

III. RESULT

Table 3 summarizes the average results for Minerva models and other models, on the evaluation datasets described in Section 2.3.

Figure 4 presents a breakdown of the MATH dataset results by subtopic. For MLU evaluations, unless otherwise noted, performance is measured using the standard 5-shot prompt per topic and picking the answer with the highest score.

These three figures show the variations in scores achieved by different models. Specifically, this figure shows how Minerva 62B performed on the National Math Exam in Poland. Interestingly enough, it managed to achieve an average score of 57% - which corresponds to the national average back in 2021.

We provide results for the latest publicly available language model from OpenAI, davinci-002, evaluated using the OpenAI API with temperature set at the organization's recommended setting (T = 0.2). The combination of training data, scale and inference techniques yields state of the art results on all the technical tasks that we considered. For all tasks, improvements are considerable over previous findings. Our main focus is on few shot evaluations but we also used Minerva to fine tune against MATH. Though we did not observe any significant improvement when doing this, when choosing a different unsupervised model - PaLM - improvements were noted for MATH specifically. This demonstrates how increasing high-quality and diverse datasets will decrease the marginal utility of standard fine-tuning methods to improve performance.

IV. ACKNOWLEDGMENT

This project is done by student Bhuvan Shridhar and Barath.M , 12th Grade students under the guidance of Associate Professor Pavan Kumar, Department of Computer Science, Kendriya Vidyalaya ASC Centre, Bangalore under Central Board of Secondary Education, Delhi, India.

V. CONCLUSION

In this paper, we discuss an approach to solving mathematical problems that rely on logic and language. We used a large language model which was trained on top of a high quality mathematical dataset to successfully perform tasks in areas such as numerical calculations and symbolic manipulations. Our method does not make use of external resources - relying only on auto regression sampling at the time of inferences. Complementary approaches which similarly rely on ways other than syntax or precise calculations include models that generate codes and those based off of formalisms; these are all but one way towards achieving our goal - an agent capable of reasoning through quantitative problems.

VI. REFERENCES

[1]. Alemi, A. A., Chollet, F., Een, N., Irving, G., Szegedy, C., and Urban, J. (2016). Deepmath - Deep Sequence Models for Premise Selection. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, pages 2243–2251, USA. Curran Associates Inc.

[2]. Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. (2021). Program synthesis with large language models. arXiv preprint arXiv:2108.07732.

[3]. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot

learners.Advances in neural information processing systems, 33:1877–1901.

[4]. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2022). Quantifying memorization across neural language models.

[5]. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. (2021). Evaluating large language models trained on code. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022).

[6]. Palm: Scaling language modeling with pathways. CKE (2021). Skale centylowe wyników - matura 2021. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems.De Moura, L. M., Kong, S., Avigad, J., van Doorn, F., and von Raumer, J. (2015). The lean theorem prover (system description). In Felty, A. P. and Middeldorp, A., editors, CADE, volume 9195 of Lecture Notes in Computer Science, pages 378–388.Springer.

Cite this article as :

Barath Maheswaran, Bhuvan Shridhar, "Is Google's New AI As Smart As A Human? ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 1, pp. 144-148 , January-February 2023.
Journal URL : https://ijsrset.com/IJSRSET229659