Data Driven Approach for Mathematical Problem Solving

Byungju Kim, Wonseok Lee, Jaehong Kim and Jungbin Im
Mathpresso Inc.
Seoul, Republic of Korea
{peyton.kim, jack.lee, julio.kim, marvin.im}@mathpresso.com

Abstract

In this paper, we investigate and introduce a novel Llama-2 based model, fine-tuned with an original dataset designed to mirror real-world mathematical challenges. The dataset was collected through a question-answering platform, incorporating solutions generated by both rule-based solver and question answering, to cover a broad spectrum of mathematical concepts and problem-solving techniques. Experimental results demonstrate significant performance improvements when the models are fine-tuned with our dataset. The results suggest that the integration of contextually rich and diverse problem sets into the training substantially enhances the problem-solving capability of language models across various mathematical domains. This study showcases the critical role of curated educational content in advancing AI research.

Keywords: Mathematical problem solving, Data-driven, Large language model

1. Introduction

In the field of machine learning, the ability to solve complex mathematical problems is often used as a measure of a model's reasoning abilities, understanding of natural language, and its capacity to engage in abstract thinking. From the basic arithmetic operations to more complex numerical challenges, the capacity of machine learning models to navigate and resolve mathematical problems lays the groundwork for advanced applications in fields such as data analysis, and education. This significance is due to the fundamental nature of mathematics as a form of structured problem-solving. The endeavor to enhance machine learning models' capability in mathematical problem-solving is driven by the dual goals of improving their analytical capabilities and enabling them to handle real-world tasks that require precise numerical computations. This pursuit involves not only refining the models' ability to understand and analyze numerical data but also their capability to interpret contextual information and properly apply mathematical concepts in varied scenarios.

On the other hand, the advent of large language models (LLMs) has marked a significant milestone in demonstrating the potential of data-driven approaches. Numerous LLMs, such as GPT (OpenAI et al., 2024), Gemini (Team et al., 2023), Llama-2 (Touvron et al., 2023; Rozière et al., 2024) and Orca-2 (Mitra et al., 2023), have demonstrated an ability to understand and generate human-like text. They have achieved exceptional performance across a variety of tasks, including mathematical problem-solving. This capability arises from their extensive training on diverse datasets, and sophisticated training algorithms to process and learn from the data. The remarkable performance of these models has shown that with sufficiently many data, it is possible to achieve levels of understanding and interpreting capabilities that closely mimic human cognitive processes.

Moreover, the scalability of large language models shows that their performance often improves with the addition of more data. This phenomenon is referred to as "scaling laws" (Johnson et al., 2018; Kaplan et al., 2020; Fernandes et al., 2023; Isik et al., 2024). This suggests that the limits of these models' capabilities are continually expanding, as more data becomes available for training.

Similar approaches have been attempted in the field of mathematical problem-solving. Llemma-2 (Azerbayev et al., 2023), part of the Llama-2 (Touvron et al., 2023) series, have been trained on a mixture of publicly available data, and achieved remarkable performances in various mathematical tasks (Hendrycks et al., 2021b; Cobbe et al., 2021a; Lewkowycz et al., 2022; Hendrycks et al., 2021a). Their performance enhancement implies that the current transformer-based (Vaswani et al., 2023) LLMs can learn mathematical induction from the large corpus of data.

Another research stream of data-driven training is to utilize several existing LLMs to synthesize new datasets (Yu et al., 2023; Toshniwal et al., 2024; Yue et al., 2023) or to teach other models (Burns et al., 2023; Luo et al., 2023). Their primary aim is to curate a variety of math problems with rich step-by-step solutions, enabling a LLM to effectively learn the logic underlying the progression of mathematical steps.

In this work, we introduce a new model based on Llama-2, trained using our dataset. Drawing inspiration from previous research, we have collected a novel dataset for math problem-solving. Our dataset collection methodology ensures that it

MathNLP: The 2nd Workshop on Mathematical Natural Language Processing @ LREC-COLING 2024, pages 25–34

mirrors the distribution of mathematical challenges encountered in the real world. This model demonstrates consistent performance improvements on math problem-solving tasks. To benchmark its performance, we conducted evaluations of our trained models against the MATH and GSM8k datasets (Hendrycks et al., 2021b; Cobbe et al., 2021b). Further analysis of our dataset's composition reveals that the observed performance improvements align with its distribution. The alignment of our dataset's distribution with the performance improvements suggests that performance could be further enhanced by expanding our data. This implies that as we enrich our dataset with a broader range of problems, the model's ability to tackle diverse mathematical tasks is likely to improve. It highlights the importance of a comprehensive dataset for optimizing performance in math problem-solving tasks.

2. Dataset Construction

To construct a dataset of math problems with explanatory solutions, we have utilized our question answering platform, [1]QANDA. Our platform serves as a math question-and-answer app, designed to bridge the gap between students facing mathematical challenges and teachers equipped to provide solutions. The interactive environment allows students to pose math problems, to which the platform's network of qualified teachers responds with detailed answers, explanations, and step-by-step guidance. Since the math problems are originated from the users who pose a question to the platform, the distribution of the problems in our dataset reflects the diverse mathematical challenges encountered by students. To secure the broader diversity in solutions, we have collected two distinct types of solutions offered by the platform: question answering and rule-based math solver.

2.1. Rule-based Math Solver

Similar to the Sympy (Meurer et al., 2017), a python library for symbolic mathematics, the platform provides a solution to a mathematical problem expressed through symbolic expressions. By adopting a rule-based approach, the solver ensures a high degree of accuracy and reliability, offering solutions that mimic the methodical process a human may use. Moreover, it offers step-by-step solutions to the given problems, thereby enabling a deeper understanding of the problem-solving process. By leveraging the platform, we have obtained detailed solutions, which demonstrate the procedural steps involved in resolving complex mathematical expressions.

[1] https://mathpresso.com/en

Problem :

Solve the equation $\dfrac{x+3}{x+2} - \dfrac{x}{x-2} = \dfrac{5x}{x^2-4}$

Solution :

If $\frac{a(x)}{b(x)} = c(x)$ is valid, it is $\begin{cases} a(x) = b(x)c(x) \\ b(x) \neq 0 \end{cases}$

$$\begin{cases} 5x = (x^2-4)\left(\dfrac{x+3}{x+2} - \dfrac{x}{x-2}\right) \\ x^2 - 4 \neq 0 \end{cases}$$

Simplify the expression

$$\begin{cases} 5x = -x - 6 \\ x^2 - 4 \neq 0 \end{cases}$$

Solve a solution to x

$$\begin{cases} x = -1 \\ x^2 - 4 \neq 0 \end{cases}$$

Substitute $x = -1$

$$\begin{cases} x = -1 \\ (-1)^2 - 4 \neq 0 \end{cases}$$

Therefore, $x = -1$

(a)

Problem :

Given two sets $X = \{-1, 0, 1\}$ and $Y = \{-2, -1, 0, 1, 2\}$, there exists a function f from X to Y. For every element x in X, when $xf(x)$ is a constant function, what is the value of $f(1) + f(-1)$?

Solution :

Given $xf(x)$ is a constant function, we have $xf(x) = c$. When $x = 0$, we have $0 \cdot f(0) = 0$. Therefore, $c = 0$. Hence, for all x, we have $xf(x) = 0$, which implies $f(x) = 0$. Therefore, $f(1) + f(-1) = 0 + 0 = 0$.

(b)

Figure 1: Data points sampled from our dataset. Given math problems, their solutions are collected from either (a) rule-based math solver and (b) question answering. Rule-base solver provides detailed step-by-step solutions, while the other covers more complex problems

The enriched datasets offer a substantial advantage in training LLMs. The detailed solutions are inherently superior for training purposes since they encompass an extensive range of information about mathematical concepts and the procedures involved in problem-solving (Lightman et al., 2023; Wang et al., 2024). Such an approach ensures the models not to estimate the mathematical reasoning but to replicate the logical deductions required to solve complex problems. Consequently, these enriched training datasets are instrumental in enhancing the capability of LLMs to solve mathematical

Training Procedure	Prealg.	Algebra	Intermediate Algebra	Number Theory	Counting & Probability	Geometry	Precalculus
				Level 1			
Llama + M	59.3	64.4	50.0	40.0	38.5	31.6	42.1
Llama + Q + M	73.3	79.3	63.5	43.3	**69.2**	55.3	61.4
Llemma + M	80.2	85.2	69.2	**63.3**	61.5	**57.9**	49.1
Llemma + Q + M	**81.4**	**87.4**	**75.0**	56.7	64.1	**57.9**	**64.9**
				Level 2			
Llama + M	48.0	49.3	17.2	21.7	25.7	30.5	15.0
Llama + Q + M	64.4	**74.6**	35.9	**39.1**	37.6	**47.6**	28.3
Llemma + M	63.3	67.7	25.0	22.8	34.7	39.0	29.2
Llemma + Q + M	**70.1**	73.1	**38.3**	37.0	**40.6**	45.1	**35.4**
				Level 3			
Llama + M	35.3	34.5	7.7	14.8	13.0	14.7	2.4
Llama + Q + M	**53.6**	55.9	14.9	**24.6**	29.0	25.5	10.2
Llemma + M	50.5	54.8	12.3	**24.6**	32.0	29.4	17.3
Llemma + Q + M	52.7	**69.0**	**21.0**	**24.6**	**36.0**	**39.2**	**21.3**
				Level 4			
Llama + M	26.7	17.3	7.3	9.2	7.2	10.4	3.5
Llama + Q + M	41.4	**49.8**	8.9	11.3	13.5	13.6	6.1
Llemma + M	38.2	34.3	7.7	**17.6**	13.5	15.2	6.1
Llemma + Q + M	**46.1**	48.1	**12.9**	**17.6**	**16.2**	**24.0**	**7.0**
				Level 5			
Llama + M	9.3	8.8	1.4	**5.8**	1.6	1.5	0.0
Llama + Q + M	17.6	21.2	2.1	5.2	5.7	1.5	1.5
Llemma + M	19.7	20.9	2.9	5.2	4.9	3.0	**4.4**
Llemma + Q + M	**21.8**	**30.6**	**3.6**	5.2	**8.1**	**3.8**	3.0
				Overall			
Llama + M	32.6	29.7	9.4	13.3	13.5	14.0	8.8
Llama + Q + M	47.1	51.3	15.1	19.1	24.5	21.9	16.3
Llemma + M	46.5	46.8	13.2	19.1	23.6	22.3	17.6
Llemma + Q + M	**50.8**	**56.9**	**18.9**	**21.1**	**27.4**	**28.0**	**21.3**

Table 1: Model performance across different mathematical domains. Models trained with our dataset show better performance in most of the mathematical domains. In the training procedure, (+Q) denotes fine-tuning with our dataset collected through QANDA, and (+M) denotes fine-tuning with Metamath dataset

problems.

Figure 1 (a) describes an exemplar data instance. In each step of the solution, the solver provides a brief explanation of the related concept before proceeding with the actual calculation.

2.2. Question Answering

Once a user pose a math problem, the platform searches from the database and curate several problem-solution pairs that match the query question. The database is constructed to aid students in understanding mathematical concepts. It makes the obtained problem and solution to be intrinsically educationally effective; the solutions are structured and detailed. These educational characteristics, such as providing step-by-step explanations and highlighting the underlying mathematical principles, benefit the training procedure of LLMs as well.

Figure 1 (b) describes an exemplar data instance collected through question answering. Comparing to data collected through rule-based math solver, data pairs gathered through question-answering mechanisms reveals a notable difference in the complexity. The problems accumulated through question answering tend to require more complex procedures to solve. It is trivial since we can easily compose a math problem with multiple expressions. This complexity demands a deeper understanding of mathematical concepts and longer deduction process. In other words, the dataset collection through question answering not only diversifies the range of problems in the dataset but also enriches it with challenges that necessitates advanced problem-solving strategies.

3. Experiments

3.1. Model Training

We fine-tuned Llama-2 7B (Touvron et al., 2023) and Llemma-2 (Azerbayev et al., 2023) 7B with our dataset. Each data instance is presented in the following prompt:

```
Problem:{math problem}
Solution:{ground truth solution}
```

To computationally evaluate performance in math problem solving, it is crucial for the model to generate an answer that is parsable. Unfortunately, as illustrated in Figure 1, achieving this is fundamentally challenging within our dataset, given that the solutions were not created in a fully controlled environment. To address this issue, we further fine-tuned our trained model using the Metamath dataset (Yu et al., 2023), enabling the model to learn the generation of well-formatted outputs. For a fair comparison, both the Llama-2 and Llemma-2 models were also trained using the Metamath dataset, utilizing the prompt described earlier.

3.2. Evaluation

To verify the effectiveness of our dataset, we evaluated the performance of MATH datasets (Hendrycks et al., 2021b). MATH is a challenging dataset designed to evaluate the mathematical problem-solving capabilities of machine learning models. It covers a wide array of domains, including prealgebrea, algebra, number theory, counting, probability, geometry, intermediate algebra, and precalculus.

Table 1 shows the performance of our approach. We break down the dataset into domains and levels to examine detailed characteristics and trends. In the training procedure, +Q denotes fine-tuning with our dataset collected through QANDA, and +M denotes fine-tuning with Metamath dataset. Note that every model undergoes fine-tuning with Metamath dataset in the end. The table distinctly demonstrates that the fine-tuning with our dataset (+Q) significantly improves the original performance, indicating a considerable improvement.

Here, it is noteworthy to focus on the difference between Llama+O and Llemma model. Llemma is initialized with CodeLlama (Rozière et al., 2024), and trained with dataset named *Proof-Pile-2* (Azerbayev et al., 2023). The dataset is composed of code (Kocetkov et al., 2022), mathematical content from web (Paster et al., 2023), and scientific papers

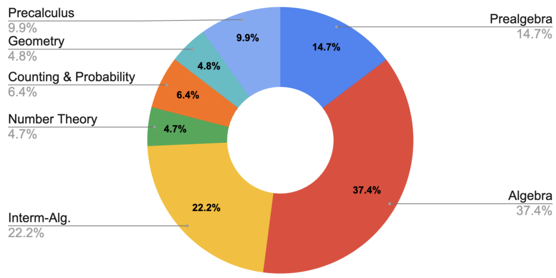

Figure 2: Domain composition of our dataset

(Computer, 2023). Our dataset, on the other hand, is fully composed of mathematical problems and their solutions.

3.3. Composition of Our Dataset

Interesting trends emerge in Algebra and Precalculus. In Algebra, the Llama+Q+M model consistently outperforms the Llemma+M model across all difficulty levels, except for the easiest set (Level 1). In contrast, within Precalculus, the Llemma+M model consistently outperforms Llama+Q+M, again with the exception of the easiest set. This suggests that the Llemma model has a stronger grasp of concepts in calculus, whereas the Llama+Q model is more adept at handling algebraic problems. Since the major difference between the two models is their training data, it is reasonable to consider that the trends implies the compositional difference between *Proof-Pile-2* and our datasets. Although not as pronounced as in the case of Precalculus, Table 1 also exhibits similar trends within the Geometry and Number Theory domains.

To explore the relationship between dataset composition and model performance, we categorized the instances in our dataset based on the domains present in the MATH dataset. Figure 2 illustrates our dataset's composition, reflecting the aforementioned trends. The majority of our dataset is classified under Algebra, Prealgebra, or Intermediate Algebra, whereas less than 10% of the samples falling into the Precalculus category. Additionally, Figure 2 shows that Geometry and Number Theory are the lesser-represented domains in our dataset. This distribution aligns with the observed performance trends.

The optimistic outlook based on this trend is that we could enhance performance in domains beyond algebra simply by collecting more data. However, this approach may potentially compromise algebra's performance.

3.4. Overall Performance

While our investigation primarily concentrated on how the composition of our dataset affects the per-

Models	GSM8k	MATH
MAmmoTH	53.6	31.5
Metamath	66.5	19.8
Llemma + M	69.2	30.0
Mistral + M	**77.7**	28.2
ToRA	68.8	**40.1**
Llama + Q + M	66.2	31.4
Llemma + Q + M	71.0	36.1

Table 2: Overall performance of various models

formance improvements within their corresponding domain, we also validated our methodology by assessing overall performance. For this experiment, we incorporated the GSM8k(Cobbe et al., 2021b) dataset. The GSM8k dataset consists of grade school math problems that require two to eight steps to solve, involving elementary-level calculations through basic arithmetic operations. For each model, we fixed their size as 7B.

Table 2 presents a comparison of the overall performance between our method and other models. It is important to note that the Metamath model is identical to Llama+M in Table 1. For both datasets, our model (Llemma+Q+M) achieves the second-highest performance. Our model notably excels in the MATH dataset over other models, with the exception of ToRA (Gou et al., 2024). Given that ToRA employs a tool-augmented, multi-step method, our findings highlight the efficacy of our data-driven approach.

4. Conclusion

This study has highlighted the potential of data-driven approaches to mimic and augment human cognitive processes in structured problem domains. By training a novel Llama-2 based model with a specially curated dataset reflective of real-world mathematical challenges, we have demonstrated significant advancements in the model's ability to tackle complex numerical tasks across a variety of mathematical domains. Our dataset, constructed from a unique blend of rule-based solutions and human-generated answers via a question-answering platform, has proven to be beneficial in achieving these improvements. The performance of our model, especially when compared against the MATH dataset, validates the efficacy of our dataset in enhancing the analytical capabilities of LLMs. The findings from this research suggest that the integration of more diverse and complex datasets would result in better performing models in mathematical domains.

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2024. Scaling laws for downstream task performance of large language models.

Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman. 2018. Predicting accuracy on large datasets from smaller pilot data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 450–455, Melbourne, Australia. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. 2017. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka,

Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol

Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters,

Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning.