Feedback or Autonomy? Analyzing LLMs' Ability to Self-Correct

Stanford CS224N Custom Project

Kai Fronsdal
Department of Computer Science
Stanford University
kaif@stanford.edu

Abstract

Self-Correction has emerged as a promising method of improving Large Language Models (LLMs) base reasoning capabilities. However, recent work has shown this approach to fail for mathematics. In this paper, we delve into the fundamental capacity of LLMs to perform self-correction, primarily in the field of mathematics. Our investigation establishes that smaller models exhibit shortcomings in accurately assessing mathematical arguments, and fine-tuning for self-correction only improves their self-calibration rather than their ability to discern correctness. Further, through a manual analysis of correct and incorrect solutions, we find that models are unable to identify granular calculation mistakes such as carry errors and missing signs, but are able to correct high-level planning. Our findings suggest self-correction appears to be an emergent behavior that smaller models fundamentally lack.

1 Key Information to include

- Mentor: Yuhui Zhang

2 Introduction

Large Language Models (LLMs) have come to dominate natural language processing in the last few years due to their powerful expressive abilities. In particular, LLMs have demonstrated exceptional stylistic and prose abilities (OpenAI et al., 2023; Chiang et al., 2023) while also showcasing surprising abilities of language understanding (Begu et al., 2023; Wei et al., 2022a). Moreover, these models are increasingly capable of completing reasoning tasks (Wei et al., 2022c; Kojima et al., 2022).

Despite the merits, LLMs are not without shortcomings. They have been observed to occasionally manifest undesired and inconsistent behaviors, such as producing convincingly inaccurate hallucinations (Zhang et al., 2023; Lin et al., 2022; Wei et al., 2022b) and promoting misleading reasoning (Golovneva et al., 2023; Wu et al., 2023). A popular method to mitigate these problems is through human feedback (Bai et al., 2022a). Human quality assessments of model outputs act as a reward signal to optimize model performance – analogous to the human learning process, which typically involves learning from mistakes via self-reflection, under the assumption that models exhibit similar behavior.

Due to the massive compute required to train more powerful models (Sastry et al., 2024) and time and cost of collecting human feedback, researchers have put significant effort into creating augmentation techniques that reduce reasoning errors such as Chain-of-Thought (Wei et al., 2022c) and Self-Consistency (Wang et al., 2022) that do not require scaling to larger and larger models. Self-correction is one promising approach where an LLM evaluates or fixes its own responses. Madaan et al. (2023) used self-correction to finetune GPT-3.5 and GPT-4 achieving an impressive average

20% improvement in their tasks. Noticeably, however, mathematical reasoning failed to improve through self-correction – a result echoed by Ye et al. (2023).

In this paper, we explore self-correction in the domain of mathematics – often used as a strong measure of model capability. In particular, our results corroborate the finding that moderate sized LLMs are poor at spotting errors in mathematical reasoning. Despite delicate finetuning designed to enhance their ability to give feedback and evaluate reasoning stages, we note no substantial improvements in their capacity to differentiate correct solutions from incorrect ones. Intrigued by this phenomenon, we scrutinized the model's outputs, leading us to discover that LLMs are considerably more adept at rectifying reasoning errors than they are at correcting algebraic or calculation-related mistakes. This provides one explanation as to why researchers have found that LLMs fare better in self-correction within domains other than mathematics.

3 Related Work

The concept of self-correction in Large Language Models has recently gained huge popularity, especially in the context of scaling models beyond human capabilities. Discussions related to self-correction focus on exploring if these highly sophisticated models have the capability of determining the accuracy of their output and refining their responses Bai et al. (2022a); Madaan et al. (2023); Lee et al. (2023); Bai et al. (2022b). To illustrate, consider a scenario where an LLM is presented with a complicated mathematical problem. It may solve it initially, but accidentally commit an error in one of the calculation stages such as a missing sign or something more significant. In an ideal situation, the model should be capable of pinpointing this potential discrepancy, review the problem, correct the error, and subsequently generate a more accurate solution.

The current literature on few-shot prompting highlights the notion of LLM self-correction, i.e., a process by which an LLM adjusts its outputs (See Pan et al. (2023) for an overview of the literature). Despite the initial promise of these approaches, Huang et al. (2023); Tyen et al. (2023) have observed that a significant proportion of the performance improvement obtained through these techniques was attributed to the use of oracle or external feedback. This involved using the ground truth to know when to prompt the model to change an answer. While this is sometimes a valid approach, such as coding problems and some computational math questions where it is easy to test for correctness Gou et al. (2023); Zhou et al. (2023), for most of informal math this is impossible. This is especially prominent in multiple choice questions, where randomly guessing multiple times with feedback results in high accuracy.

Interestingly, evidence from Tyen et al. (2023) demonstrates that for various tasks, LLMs are capable of correcting their reasoning errors when clearly pointed out, but struggle to identify these errors independently. Additionally, Wang et al. (2023) observed that models can be easily swayed through biased feedback. Thus, our study primarily emphasizes evaluating models' capacity for error detection, as opposed to their correction abilities.

Several iterative approaches resolve some of these issues. Yao et al. (2023); Xie et al. (2023) both found that adding self-reflection at each reasoning step dramatically improved performance for reasoning tasks. This suggests that one of the challenges in self-correction, particularly in mathematical reasoning, is akin to locating a needle in a haystack. Assessing the correctness of an entire statement in one go proves significantly more difficult than evaluating it in smaller segments.

4 Approach

We begin by evaluating the capabilities of LLaMA-2 with 7B parameters (Touvron et al., 2023). We perform a systematic evaluation of prompting methods on self-correction ability. In particular, how often does it rate a solution as correct given it is correct and how often it rates a solution as incorrect given it actually is is incorrect. Notably, we report similar findings to Huang et al. (2023); Tyen et al. (2023).

To examine if finetuning the model can enhance its performance, we adopt Low-Rank Adaptation (LoRA) due to computational constraints. LoRA is a well-known method for efficiently training large-scale models (Hu et al., 2022a). Rather than finetuning the entire model, the original weight matrix of the pretrained model remains frozen and only significantly smaller row rank matrices

Subject	Accuracy
Algebra	0.63
Counting and Probability	0.31
Geometry	0.22
Intermediate Algebra	0.23
Number Theory	0.48
Prealgebra	0.65
Precalculus	0.39
GSM8K	0.86

Table 1: Accuracy of the DeepSeekMath-Instruct 7B model using 4-shot prompting. All rows but GSM8K are different splits of the MATH dataset. The model has an overall average accuracy of 47%, so we get a fairly even split of correct and incorrect solutions.

undergo updates. This significantly reduces the quantity of trainable parameters, thereby dramatically decreasing memory consumption and training time. We also experimented with prompt tuning, which given the model's high sensitivity to prompts seems like a reasonable experiment. However, we encountered technical difficulties that we were unable to address within our timeline.

We conclude our study with a qualitative examination of the model's outputs, primarily scrutinizing the specific error trends that the model can correct and those it tends to repeat. Considering the model's suboptimal performance, we hypothesis that self-correction is likely an emergent behavior.

We used the HuggingFace libraries transformers, peft, and accelerate to simplify finetuning the model but wrote the dataloader and training loop from scratch. We also used the library vllm to speed up inference of the models. Everything else we programmed from scratch.

5 Experiments

5.1 Data

GSM8K (Cobbe et al., 2021) is a dataset of high-quality linguistically diverse grade school math word problems created by human problem writers. The solutions primarily involve performing a sequence of elementary calculations.

Hendrycks **MATH** (Hendrycks et al., 2021) is a dataset of 12,500 challenging competition mathematics problem across seven different areas of high school math: algebra, counting and probability, geometry, intermediate algebra, number theory, prealgebra, and precalculus. The solutions require multiple correct non-trivial reasoning steps to get right as well as some hard calculations.

Both datasets have simple final answers such as a single number or a multiple-choice option. When evaluating models on these benchmarks, we instruct them to reason about the problem before providing their final solution in a boxed format (i.e. `\boxed{...}`) which is easy to automatically parse. Post-processing is then employed to account for alternative correct LaTeX representations. For instance `0.5`, `1/2`, `\frac{1}{2}` and `\dfrac{1}{2}` should all be recognized as equivalent solutions.

For evaluating how capable our model is at evaluating correctness, due to the poor base performance of LLaMa-2 7B on MATH and GSM8K, we generated solutions using DeepSeekMath-Instruct 7b (Shao et al., 2024), a strong model with good mathematical capabilities for a model of this size. See Table 1 for the distribution of correct and incorrect answers we generated. Even though this is no longer true self-correction, we believe this modification is satisfactory enough for the purposes of this analysis.

For finetuning, we employed the SelFee feedback dataset (Ye et al., 2023) – a collection of instruction-answer pairs enriched with chatGPT (OpenAI, 2022) feedback and iterative revisions. This dataset includes math and code tasks as well as general instruction and chat data to improve generalization. We also generated and filtered by hand a dataset of 120 examples of feedback in the format of our specific questions (split evening between correct and incorrect answers as well as three different kinds of feedback described in section 5.3.1).

5.2 Evaluation method

We evaluate our baseline model's reasoning ability using overall accuracy, extracting the boxed answer from the generated solution and comparing it against the ground truth. To evaluate self-correction ability we compare the distributions of feedback against solution correctness. To get a quantitative result, we then perform a MannWhitney U test (Mann and Whitney, 1947) to compare the distribution of confidences conditioned on correctness. We repeat this process for the finetuned model.

5.3 Experimental details

5.3.1 Collecting Feedback

We tested three different categories of feedback. For each category due to the sensitivity of LLMs to the prompt, we tested several different variants. For a full list of tested prompts, see the appendix. We generally asked the model to explain its reasoning before giving a final rating.

1. **Confidence**: We asked the model to evaluate it's confidence the given solution was correct.
2. **Mistake**: We asked the model to find mistakes in the solutions.
3. **Check**: We asked the model to double check all of the calculations.

Additionally, we tried several different rating scales

1. **Percent**: the models rates its confidence in the correctness of the solution on a scale from 0 to 100%
2. **Rating**: the model rates the accuracy of the solution on a scale from 1-5
3. **Correctness**: the model outputs "correct" or "incorrect."
4. **Confidence Level**: the model outputs its confidence as one of "very confident," "confident", "somewhat confident", "not very confident", or "not confident"

We found that there was not a massive difference in performance between the four different scales. It is important to note that we rescaled each of the rating scales to be between 0 and 100 so it was easier to compare between the prompts.

For each prompt and dataset pair, we evaluated the models on a random sample of 100 question-solution pairs.

5.3.2 Finetuning

We ran experiments using LLaMA-2 with 7B parameters (Touvron et al., 2023). We finetuned on the SelFee dataset with LoRA for 20k steps with the AdamW optimizer (Loshchilov and Hutter, 2017) using a learning rate of 3e-4 and batch size of 4. For LoRA, we set the hyperparameters to be $\alpha = 15$, 10% dropout, rank 64, and no bias similar to Hu et al. (2022b).

Due to limited compute availability, we were unable to perform much hyperparameter optimization, which could be a reason for the lack of improvement on our self-correction task.

5.4 Results

In Figure 1 we see that the LLM exhibits excessive overconfident in the solutions irrespective of the prompt. On the surface there seems to be a small improvement after finetuning; we see in Figure 2 that the distribution of responses is better balanced indicating a better-calibrated model. However, when we ran a Mann-Whitney U test (Table 2 and 3), comparing the distributions of confidences, no statistically significant difference was evident in the responses from both the baseline and finetuned models – regardless of the correctness of the provided solution.

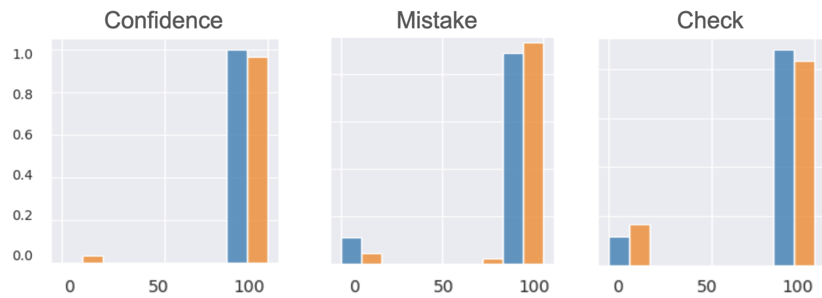

Figure 1: Distributions of baseline model's confidence conditioned on solution correctness. Confidences of correct solutions are given in orange and confidences of incorrect solutions in blue. Listed here are the best best cherry picked prompts for each category, averaged over each dataset. For the prompts used and the more finegrained plots, see the appendix.

Figure 2: Distributions of finetuned model's confidence conditioned on solution correctness. Confidences of correct solutions are given in orange and confidences of incorrect solutions in blue. Listed here are the best best cherry picked prompts for each category, averaged over each dataset. For the prompts used and the more finegrained plots, see the appendix.

Prompt Idea	Algebra	Count. & Prob.	GSM8K	Geometry	Inter. Alg.	Num. Theory	Prealg.	Precalc.
Confidence	0.67	0.32	0.82	0.40	0.75	0.32	0.19	0.45
Mistake	0.40	0.13	0.29	0.75	0.74	0.86	0.40	0.31
Check	0.43	0.30	0.30	0.30	0.25	0.56	0.47	0.23

Table 2: p-values for the Mann-Whitney U tests comparing the distributions of confidences for correct and incorrect solutions on the baseline model. Notice that all of them are well above a reasonable 0.05 significance level.

Prompt Idea	Algebra	Count. & Prob.	GSM8K	Geometry	Inter. Alg.	Num. Theory	Prealg.	Precalc.
Confidence	0.85	0.43	0.20	0.48	0.66	0.31	0.21	0.37
Mistake	0.24	0.28	0.30	0.72	0.88	0.60	0.37	0.30
Check	0.38	0.11	0.28	0.30	0.29	0.53	0.41	0.25

Table 3: p-values for the Mann-Whitney U tests comparing the distributions of confidences for correct and incorrect solutions on the finetuned model. The average p-value is about 0.037 lower after finetuning. Each result is still above a reasonable 0.05 significance level.

While the p-values tend to be a bit lower on average after fine-tuning, this is not substantial evidence to base a claim on [1].

We conclude that smaller LLMs are not capable of reliable self-correction without external feedback. Similar to other emergent behaviors, such as Chain-of-Thought (Wei et al., 2022c), it appears that smaller models do not have the ability to detect subtle errors in reasoning or calculation mistakes.

6 Analysis

In order to investigate the cause of the our model's failure to catch mistakes in mathematical reasoning, we examined a random sample of 120 reflections. We split mistakes into two categories: reasoning and calculation. There are many nuances to mistakes, but for the sake of brevity we decided on these two. The evaluation is somewhat subjective, but still enlightening nonetheless. Reasoning mistakes included using incorrect formulas, the wrong high-level approach, copying incorrectly, and hallucinations. Calculation mistakes included small typos, incorrect algebraic manipulation, sign errors, and wrong numerical calculations.

In Table 4 we find that the model is much worse at correcting calculation errors than reasoning errors. In particular, when the ground truth answer was correct, the model is more likely to falsely identify a calculation mistake (40%) than a reasoning mistake (20%). When the ground truth answer was incorrect, the model is more likely to catch a reasoning error (65%) than a calculation error (30%). We had expected asking the model to double check the calculations to improve the models ability to find calculation errors, but we did not find this in practice.

We also found that the model was incredibly fickle. It would often change its mind when sampling feedback for the same response. This lends credence to the claim that mathematical reasoning and in particular the ability to accurately perform calculations is an emergent behavior, but requires more investigation.

7 Conclusion

In this paper we examined the fundamental ability of small-scale LLMs to engage in self-correction. It became clear that these models fail to demonstrate accurate judgment in relation to mathematical arguments, most notably when it comes to verifying calculations. Although fine-tuning focused on enhancing self-correction capabilities did manage to improve the model's self-calibration (thus curbing overconfidence), it fell short of elevating its proficiency in accurately discerning correctness. While we acknowledge that our research examined only a single model with limited hyperparameter tuning, and it is feasible that enhanced optimization or alternate models could yield better results, we claim that self-correction is an emergent behavior, a trait smaller models smaller models fundamentally lack the capacity to exhibit.

[1] Even if a few of the p-values appeared significant, there is a high likelihood it would be a result of p-hacking. This could be mitigated through multiple-hypothesis testing methods, but because we are nowhere close to rejecting the null hypothesis, this is not an issue.

	Error Type		
Ground Truth	None	Reasoning	Calculation
Correct	0.4	0.2	0.4
Incorrect	0.7	0.65	0.35

Table 4: Proportion of error modes during self-correction. For the correct ground truth answers, each entry corresponds to the proportion of responses where the model made the given error type when evaluating correctness (this row sums to 1). The model made fewer reasoning errors when evaluating solutions than calculation errors. For the incorrect ground truth answers, each entry corresponds to the proportion of the given error type the model successfully corrects. The model had a higher chance of correcting a reasoning mistake than a calculation one.

One potential avenue for future research could delve deeper into more specific error types that these models produce to ascertain if they can be mitigated. Additionally, it is worth noting that Yao et al. (2023); Xie et al. (2023) found that assessing each step of a response individually yielded better results than evaluating the entire response collectively. It would be interesting to see which types of errors this approach can prevent. We would also welcome future experiments involving larger models to determine if our hypothesis on self-correction as an emergent behavior holds true.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv: 2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv: 2212.08073*.

Gaper Begu, Maksymilian Dbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv: 2305.00948*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

O. Yu. Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. *International Conference on Learning Representations*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv: 2305.11738*.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022b. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv: 2310.01798*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv: 2309.00267*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv: 1711.05101*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *NEURIPS*.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Blog post.

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, ukasz 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, ukasz 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. 2023. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv: 2308.03188*.

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. 2024. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv: 2402.08797*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

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. *arXiv preprint arXiv: 2307.09288*.

Gladys Tyen, Hassan Mansoor, Victor Crbune, Peter Chen, and Tony Mak. 2023. Llms cannot find reasoning errors, but can correct them! *arXiv preprint arXiv: 2311.08516*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *Conference on Empirical Methods in Natural Language Processing*.

Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, E. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv: 2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv: 2307.02477*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv: 2305.10601*.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv: 2305.13534*.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv: 2308.07921*.

A Appendix (optional)

A.1 Prompts Tested

```
confidence_prompt = [
    "How confident are you this is correct? Explain your reasoning. Give a
    1-5 rating, 1 being not confident at all and 5 being very confident. End
    your response with 'Rating: ' followed by your rating.",
    "Are you sure this is correct? Give your confidence as a percentage
    from 0 to 100.\n\nConfidence: ",
    "Are you sure this is correct? Explain your reasoning and give your
    confidence as a percentage from 0% to 100%.",
    "How confident are you that this proof is right? Explain your
```

Prompt Idea	Algebra	Count. & Prob.	GSM8K	Geometry	Inter. Alg.	Num. Theory	Prealg.	Precalc.
Rating	0.67	0.32	0.82	0.40	0.75	0.32	0.19	0.45
Percent	0.40	0.13	0.29	0.75	0.74	0.86	0.40	0.31
Percent	0.82	0.95	0.35	0.08	0.96	0.16	0.59	0.67
Confidence Level	0.56	0.75	0.16	0.17	0.56	0.75	0.73	0.20
Percent	0.43	0.11	0.30	0.30	0.84	0.56	0.47	0.23
Percent	0.59	0.74	0.52	0.17	0.88	0.21	0.89	0.61
Correctness	0.69	0.26	0.93	0.14	0.85	0.53	0.51	0.68

Table 5: *p*-values for the Mann-Whitney U tests comparing the distributions of confidences for correct and incorrect solutions on the baseline model. Notice that all of them are well above a reasonable 0.05 significance level.

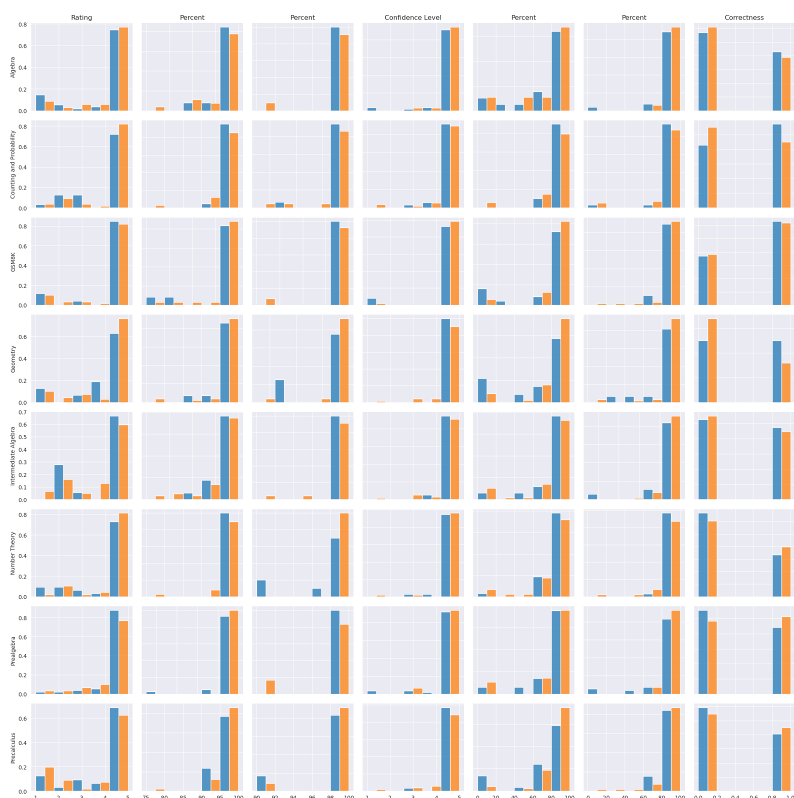

Figure 3: All finetuned model confidence predictions.

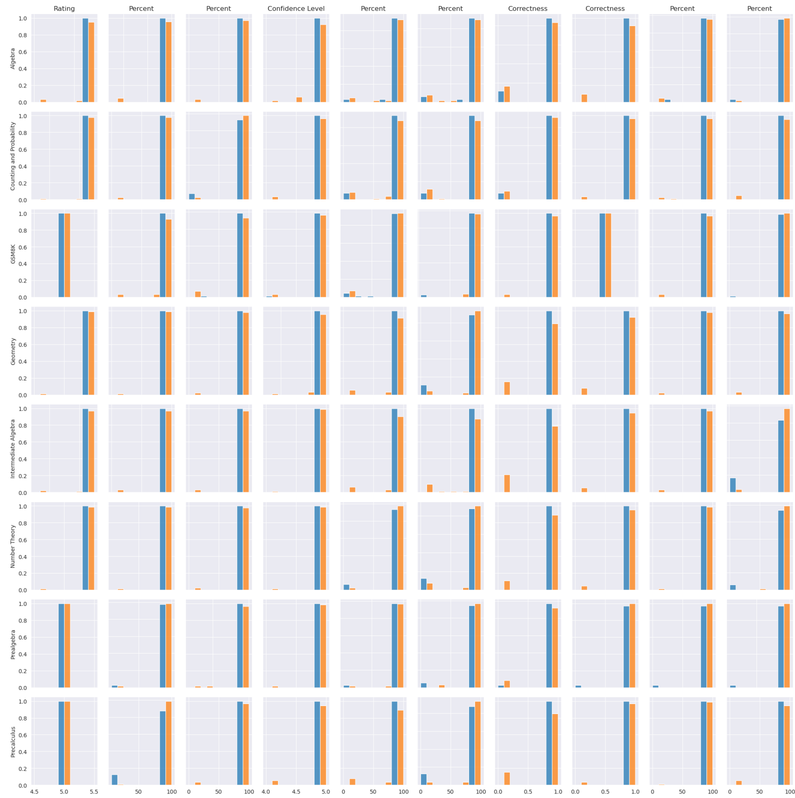

Figure 4: All base model confidence predictions.

```
    reasoning and say whether you are very confident, confident, somewhat
    confident, not very confident, or not confident.",
    "Above is a student response to a question. Pretend you are a teacher
    grading the response. How confident are you that the student's response
    is correct? Explain your reasoning and give your confidence as a
    percentage from 0% to 100%.",
    "Above is a student response to a question on a test. You are an
    expert and nuanced grader who has a great intuition for these things.
    How confident are you that the student's response is correct? Explain
    your reasoning and give your confidence as a percentage from 0% to
    100%.",
    "Above is a student response to a question on a test. Are there any
    mistakes in the student's response? If so, explain the mistake and say
    the student's response is incorrect. If not, say the student's response
    is correct.",
    "Are there any mistakes in this answer? If so, explain the mistake and
    say the student's response is incorrect. If not, say the student's
    response is correct.",
    "Are you sure this is correct? Explain your reasoning and give your
    confidence as a percentage from 0% to 100%.",
    "How confident are you that there is a mistake in this answer? Explain
    your reasoning and give your confidence as a percentage from 0% to
    100%.",
]

fix_mistake_prompt = [
    ["There is a mistake in this response. Explain the mistake.",
     "Craft a response that doesn't make this mistake. Do not apologize for
```

```
     the mistake or make any similar comment. Just create a new response
     that doesn't make the mistake that can replace the old response."],
    ["What is wrong with this answer?",
     "Now construct a response that fixes this mistake. Do not apologize for
     the mistake. Simply answer the question correctly without making any
     other comments."]
]

check_prompt = [
    ["Double check each of the calculations to make sure they are correct.",
    "Is there a mistake? Answer Yes or No."],
    ["We think there might be an error in the calculations about. Can you
    double check all of them by hand.", "Is there a mistake? Answer Yes or
    No."],
    ["Go over each of the calculations in the response very carefully. Can
    you find any mistakes", "Is there a mistake? Answer Yes or No."]
]
```