MAX IT OR MISS IT: BENCHMARKING LLM ON SOLVING EXTREMAL PROBLEMS

Binxin Gao
University of Maryland
bgao666@umd.edu

Jingjun Han
Fudan University
hanjingjun@fudan.edu.cn

ABSTRACT

Test-time scaling has enabled Large Language Models (LLMs) with remarkable reasoning capabilities, particularly in mathematical domains, through intermediate chain-of-thought (CoT) reasoning before generating final answers. However, the specific sources and mechanisms underlying these reasoning capabilities remain insufficiently understood. Optimization reasoning, *i.e.* finding extrema under constraints, represents a fundamental abstraction that underpins critical applications in planning, control, resource allocation, and prompt search. To systematically evaluate this capability, we introduce **`ExtremBench`**, a benchmark dataset for solving mathematical extremal problems, curated from inequality exercises used for Chinese Mathematical Olympiad and transformed into 93 standardized extrema-finding problems. We conduct extensive evaluations across various state-of-the-art open-source model families, including the Qwen3, GPT-OSS, and DeepSeek. Our results reveal that LLMs' extremal-solving reasoning capabilities do not always align with those of current mathematical benchmarks such as AIME25, with some models showing strong general mathematical reasoning but poor extremal-solving skills, and vice versa. This discrepancy highlights a critical gap in current evaluation practices and suggests that existing benchmarks may not comprehensively capture the full spectrum of mathematical reasoning abilities. Our code and dataset will be available upon acceptance.

1 INTRODUCTION

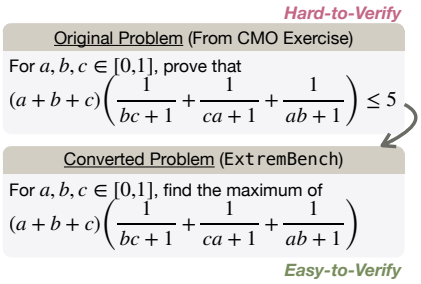

Figure 1: `ExtremBench` turns proof-style inequalities into equivalent extrema tasks, preserving reasoning challenge but enabling automatic evaluation.

Test-time scaling has enabled Large Language Models (LLMs) with remarkable reasoning capabilities, particularly in mathematical domains, through intermediate chain-of-thought (CoT) reasoning before generating final answers (DeepSeek-AI et al., 2025; OpenAI et al., 2024; Snell et al., 2024; Yang et al., 2025). However, the specific sources and mechanisms underlying these reasoning capabilities remain insufficiently understood. Optimization reasoning, *i.e.*, finding extrema under constraints, represents a fundamental abstraction that underpins critical applications in planning, control, resource allocation, and prompt search (Wang et al., 2024a; Zou et al., 2025). From determining optimal hyperparameters in machine learning (Jin et al., 2024) to solving economic equilibrium problems (Karten et al., 2025), the ability to identify extrema under given constraints forms a cornerstone of mathematical problem-solving.

Despite the fundamental importance of extremal-solving capabilities, current mathematical benchmarks such as GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021), and AIME have barely evaluated this specific form of reasoning. These benchmarks primarily focus on algebraic manipulation and arithmetic computation, leaving optimization reasoning largely unexplored. This gap is particularly concerning given that extremal problems require a distinct set of reasoning skills including identifying constraint boundaries, understanding trade-offs between competing objectives, and recognizing when optimal solutions occur at critical points or boundaries.

Can we create a math benchmark that specifically evaluates LLM's extremal reasoning capabilities? To systematically evaluate it, we introduce `ExtremBench`, a benchmark dataset for solving mathematical extremal problems, curated from inequality exercises used in Chinese Mathematical Olympiad and transformed into standardized extrema-finding tasks. Specifically, our transformation proces converts problems like "*For positive reals* a, b, c *with* $a+b+c = 3$*, prove that* $\frac{1}{a}+\frac{1}{b}+\frac{1}{c} \geq 3$" into "*Find the minimum value of* $\frac{1}{a} + \frac{1}{b} + \frac{1}{c}$ *subject to* $a, b, c > 0$ *and* $a + b + c = 3$." This reformulation, as demonstrated in Figure 1, preserves the mathematical complexity and reasoning requirements while creating a standardized and easy-to-verify [1] format for evaluating optimization capabilities.

We conduct extensive evaluations across various state-of-the-art open-source reasoning LLM families, including Qwen3 (Yang et al., 2025), GPT-OSS (OpenAI et al., 2025), and DeepSeek (DeepSeek-AI et al., 2025). Our results reveal that LLMs' extremal-solving reasoning capabilities do not always align with their performance on current mathematical benchmarks such as AIME25 and MATH-500, with some models showing strong general mathematical reasoning but poor extremal-solving skills, and vice versa. This discrepancy highlights a critical gap in current evaluation practices and suggests that existing benchmarks may not comprehensively capture the full spectrum of mathematical reasoning abilities.

Our contributions and findings are summarized as follows: ***(i)*** We construct `ExtremBench`, a dataset of 93 extremal problems from inequality proof problems used in Chinese Mathematical Olympiad. While current RL training relies heavily on verifiable answers, we introduce a perspective where we can convert hard-to-verify math proofs into numerically verifiable format, enabling systematic evaluation of optimization reasoning. ***(ii)*** We evaluate diverse state-of-the-art reasoning LLMs across multiple model families and scales, providing the first comprehensive assessment of extremal-solving capabilities in contemporary language models. ***(iii)*** Our results demonstrate that LLMs good at general math benchmarks do not always perform well at solving extremal problems, calling for more domain-specific benchmarks to evaluate LLMs' specific mathematical reasoning capabilities.

2 Related Works

Mathematical Benchmark Datasets. The evaluation of mathematical reasoning capabilities in LLMs has been facilitated by a diverse landscape of benchmark datasets spanning various difficulty levels and mathematical domains. Foundational datasets like GSM8K (Cobbe et al., 2021) and MATH-500 (Hendrycks et al., 2021) established early standards with grade-school and high school competition problems respectively, while MATHQA (Amini et al., 2019) provides GRE/GMAT-level multiple-choice questions. For formal theorem proving, miniF2F (Zheng et al., 2021) offers a cross-system benchmark with problems from mathematical olympiads. More specialized benchmarks target specific mathematical areas. FIMO (Liu et al., 2023) focuses on IMO-level algebra and number theory, PutnamBench (Tsoukalas et al., 2024) features problems from the prestigious Putnam competition, and ProofNet (Azerbayev et al., 2023) addresses undergraduate-level mathematics autoformalization. Recent efforts have pushed toward more challenging problems, with JEEBENCH (Arora et al., 2023) covering college-level topics including ODEs and multivariable calculus, MATHBENCH (Liu et al., 2024) providing hierarchical coverage from elementary to university mathematics, and GHOSTS (Frieder et al., 2024) including graduate-level exercises from advanced mathematics textbooks. To address gaps in specific domains, specialized benchmarks have emerged. CombiBench (Liu et al., 2025) provides the first comprehensive benchmark for combinatorial mathematics in Lean4 with 100 problems spanning from middle school to IMO level, while HARDMATH (Fan et al., 2024) uniquely targets asymptotic analysis and approximation methods with generated problems requiring dominant balance techniques.

LLM Reasoning for Math The capacity of LLMs to perform mathematical reasoning was significantly unlocked by CoT prompting, which elicits intermediate reasoning steps to guide the model toward a solution (OpenAI et al., 2024; Wei et al., 2022; Yang et al., 2025). Subsequent work enhanced this process through techniques like self-consistency (Wang et al., 2022), which samples multiple reasoning paths and selects the most frequent answer, mitigating errors from greedy decod-

[1] We use https://github.com/huggingface/Math-Verify for answer verification.

ing. Progress in the field has been largely driven and measured by performance on a hierarchy of benchmarks, from grade-school word problems such as GSM8K (Cobbe et al., 2021) to challenging competition-level mathematics found in the MATH-500 dataset (Hendrycks et al., 2021) and the American Invitational Mathematics Examination (AIME). A burgeoning area of research now applies LLMs to optimization, primarily focusing on translating natural language problem descriptions into formal models for external solvers, a task evaluated by benchmarks like OptiBench (Wang et al., 2024b) and addressed by frameworks such as LLMOPT (Jiang et al., 2024). However, this paradigm of problem *formulation* for numerical solvers is distinct from the symbolic and logical reasoning required to *solve* mathematical optimization problems directly. As our work highlights, existing benchmarks, while broad, do not systematically evaluate an LLM's intrinsic ability to find extrema under constraints through symbolic manipulation. This leaves a critical gap in understanding a fundamental component of mathematical intelligence, which our proposed `ExtremBench` aims to address.

3 The ExtremBench Dataset

Our dataset construction converts inequality proof problems from mathematical olympiad exercises into standardized extremal problems. We use Anthropic's Claude Opus 4.1 to execute our construction. We describe our methodology below.

Source Material. We sourced our problems from *An Introduction to the Proving of Elementary Inequalities* (Han, 2011), a popular collection of CMO exercises that contains challenging inequality problems. Given that the original problems were in Chinese, we first employed an LLM to translate them into English while preserving mathematical notation and logical structure. This translation step was verified by bilingual authors to ensure mathematical accuracy and clarity were maintained.

Transformation to Extremal Problems. For each inequality problem of the form "prove that $A \leq B$" or "prove that $A \geq B$" under given constraints, we reformulated it as an optimization task: "find the maximum/minimum of $A - B$" subject to the same constraints. Specifically, we employed the following prompt for automated conversion:

> Rewrite the inequality proof problem $A \leq B$ or $A \geq B$ as an extremum problem: with all original conditions unchanged, reformulate it as "find the extremum of $A - B$." Move all constant terms to the other side so that the objective $A - B$ includes only variable-dependent terms. Output **only** the converted problem inside a single LaTeX code block, with no additional text or explanations.

Each converted problem passes rigorous manual verification by the authors to ensure: (1) the transformation correctly preserved all constraints from the original problem, (2) the extremal formulation was mathematically equivalent to the original inequality, and (3) the problem statement was unambiguous and self-contained. From the initial 100 problems, we retained 93 after filtering out strict inequalities where equality cannot be achieved.

The Final Dataset. Our final `ExtremBench` dataset comprises 93 extremal problems, each formulated as a well-defined optimization task asking for either a maximum or minimum value, along with its corresponding numerical answer. Each problem in the dataset includes clear constraint specifications and an objective function exactly derived from the original inequality. Our dataset includes 62 minimization problems and 31 maximization problems. Examples of both the original inequality problems and their converted extremal counterparts are illustrated in Figure 1 and Appendix A.

4 Experiments

Experimental Setup. We evaluate a set of state-of-the-art open-source reasoning LLMs across three model families to assess their extremal-solving capabilities on `ExtremBench`, including Qwen3 (1.7B, 4B-Thinking-2507, 8B, 14B, 32B, 30B-A3B-Thinking-2507, 235B-A22B-Thinking-2507), GPT-OSS (20B, 120B), and DeepSeek-R1 (1.5B, 7B, 8B, 14B). We use SGLang Zheng et al.

(2024) on NVIDIA B200 GPUs for efficient inference. We report average performance of 3 repeated trials.

Benchmark Results. Figure 2 reveals surprising disparities between extremal-solving and general mathematical reasoning capabilities across model families. We draw several findings: (1) While GPT-OSS-120B-High and GPT-OSS-20B-High achieve near-perfect scores on AIME25 ($> 90\%$), their `ExtremBench` performance plateaus around 70%, suggesting that strong general mathematical reasoning does not guarantee proficiency in optimization tasks. (2) Larger models do not consistently outperform smaller ones on `ExtremBench`, *e.g.* Qwen3-14B matches the performance of Qwen3-235B despite having $17\times$ fewer parameters, indicating that extremal-solving ability may depend more on specific training data or architectural choices than raw scale. (3) The Qwen3-Thinking variants demonstrate the strongest `ExtremBench` performance ($75 - 80\%$) despite moderate AIME25 scores, while DeepSeek-R1 models show consistently lower performance on both benchmarks ($50 - 60\%$ on `ExtremBench`, $30 - 40\%$ on AIME25). These results underscore that extremal problem-solving represents a distinct mathematical competency that current benchmarks fail to capture, highlighting the necessity of specialized evaluation frameworks like `ExtremBench` for comprehensive assessment of LLM mathematical capabilities.

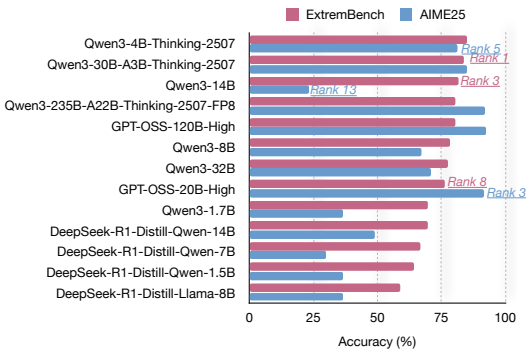

Figure 2: Evaluation results on `ExtremBench` and AIME25, sorted by the accuracy of `ExtremBench`.

5 Conclusion

In this work, we introduced `ExtremBench`, a specialized benchmark for evaluating LLMs' extremal-solving capabilities, addressing a critical gap in the benchmark of mathematical reasoning abilities. By transforming inequality proof problems from Chinese Mathematical Olympiad exercise into standardized, numerically verifiable extrema-finding tasks, we provide the first systematic evaluation framework for extremal optimization reasoning, which is a fundamental skill underlying applications in planning, control, resource allocation, and prompt search. Our extensive evaluation across state-of-the-art model families including Qwen3, GPT-OSS, and DeepSeek reveals a suprising insight: performance on extremal problems does not necessarily correlate with success on established mathematical benchmarks like AIME25 and MATH-500, with some models excelling at general mathematical reasoning while struggling with optimization tasks, and others showing the opposite pattern. This discrepancy exposes a significant blind spot in current evaluation practices and suggests that existing benchmarks fail to capture the full spectrum of mathematical intelligence. Our findings call for a more nuanced approach to benchmarking LLM capabilities, emphasizing the need for domain-specific evaluations that can identify distinct reasoning competencies beyond general mathematical problem-solving.

Future work could explore several directions: (1) Our transformation methodology from hard-to-verify inequality proofs to numerically verifiable extremal problems demonstrates a novel paradigm that could be extended to other mathematical domains. This approach could potentially unlock vast repositories of proof-based problems for RL training, as verifiable answers are crucial for test-time scaling and reward modeling. We envision similar transformations for problems in combinatorics (converting existence proofs to counting problems), geometry (converting congruence proofs to distance optimization), and analysis (converting convergence proofs to rate-of-convergence optimization). (ii) Expanding ExtremBench to include multi-objective optimization, constrained optimization with equality constraints, and discrete optimization problems would provide a more comprehensive evaluation of LLMs' optimization reasoning capabilities across different mathematical structures. (2) Investigating the underlying mechanisms behind the observed discrepancy between general mathematical reasoning and extremal-solving abilities could reveal fundamental insights about how LLMs encode and apply different types of mathematical knowledge, potentially inform-

ing more targeted training strategies. (3) Developing specialized supervised fine-tuning or RL approaches that specifically target extremal reasoning, possibly through curriculum learning that progressively increases constraint complexity and dimensionality. (4) Exploring whether the extremal-solving capability serves as a better predictor for downstream tasks in scientific computing, operations research, and automated theorem proving, where optimization reasoning is fundamental.

ACKNOWLEDGEMENT

We thank Pingzhi Li for thoughtful discussions and feedback during this work.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of NAACL-HLT*, pp. 2357–2367, 2019.

Daman Arora, Himanshu Singh, et al. Have llms advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7527–7543, 2023.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI, Daya Guo, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*, 2024.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36, 2024.

JJ Han. An introduction to the proving of elementary inequalities. *Harbin Institute of Technology Press, Harbin*, 2011.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

Caigao Jiang, Xiang Shu, Hong Qian, Xingyu Lu, Jun Zhou, Aimin Zhou, and Yang Yu. LLMOPT: Learning to define and solve general optimization problems from scratch. *arXiv preprint arXiv:2410.13213*, 2024.

Xiaolong Jin, Kai Wang, Dongwen Tang, Wangbo Zhao, Yukun Zhou, Junshu Tang, and Yang You. Conditional lora parameter generation, 2024. URL https://arxiv.org/abs/2408.01415.

Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. Llm economist: Large population models and mechanism design in multi-agent generative simulacra. *arXiv preprint arXiv:2507.15815*, 2025.

Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju, Chuanyang Zheng, Yutao Yin, Lin Li, et al. Fimo: A challenge formal dataset for automated theorem proving. *arXiv preprint arXiv:2309.04295*, 2023.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024.

Junqi Liu, Xiaohan Lin, Jonas Bayer, Yael Dillies, Weijie Jiang, Xiaodan Liang, Roman Soletskyi, Haiming Wang, Yunzhou Xie, Beibei Xiong, Zhengfeng Yang, Jujian Zhang, Lihong Zhi, Jia Li, and Zhengying Liu. Combibench: Benchmarking llm capability for combinatorial mathematics. *arXiv preprint arXiv:2505.03171*, 2025.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec

Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *arXiv preprint arXiv:2407.11214*, 2024.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024a. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL http://dx.doi.org/10.1007/s11704-024-40231-1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhuohan Wang, Ziwei Zhu, Yizhou Han, Yufeng Lin, Zhihang Lin, Ruoyu Sun, and Tian Ding. OptiBench: Benchmarking large language models in optimization modeling with equivalence-detection evaluation. *arXiv preprint arXiv:2409.16709*, 2024b.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Dongyuan Li, Renhe Jiang, Xue Liu, and Philip S. Yu. Llm-based human-agent collaboration and interaction systems: A survey, 2025. URL https://arxiv.org/abs/2505.00753.

A EXAMPLES IN EXTREMBENCH

Below we present representative examples among 93 problems from `ExtremBench`:

Example 1: For $a, b, c, d \in \mathbb{R}^+$ with $abcd = 1$, find the minimum of

$$\frac{1}{(1+a)^2} + \frac{1}{(1+b)^2} + \frac{1}{(1+c)^2} + \frac{1}{(1+d)^2}$$

Answer: 1

Example 2: For $a, b, c, d > 0$ with $a + b + c + d = 1$, find the minimum of

$$(1-\sqrt{a})(1-\sqrt{b})(1-\sqrt{c})(1-\sqrt{d}) - \sqrt{abcd}$$

Answer: 0

Example 3: For $a, b, c > 0$ with $a + b + c = 3$, find the minimum of

$$\sqrt{3-bc} + \sqrt{3-ca} + \sqrt{3-ab}$$

Answer: $3\sqrt{2}$

Example 4: For $a, b, c \in [0, 1]$, find the maximum of

$$(a+b+c)\left(\frac{1}{bc+1} + \frac{1}{ca+1} + \frac{1}{ab+1}\right)$$

Answer: 5

Example 5: For $x, y, z > 0$, find the minimum of

$$\sum_{cyc} \sqrt{\frac{(y+z)^2 yz}{(x+y)(x+z)}} - \sum x$$

Answer: 0