

MMRefine: Unveiling the Obstacles to Robust Refinement in Multimodal Large Language Models

Gio Paik*
Theta One, Inc.
giopaik@thetaone.co

Geewook Kim
NAVER Cloud AI
KAIST AI
gwkim.rsrch@gmail.com

Jinbae Im†
NAVER Cloud AI
jinbae.im@navercorp.com

Abstract

This paper introduces MMRefine, a Multi-Modal Refinement benchmark designed to evaluate the error refinement capabilities of Multi-modal Large Language Models (MLLMs). As the emphasis shifts toward enhancing reasoning during inference, MMRefine provides a framework that evaluates MLLMs’ abilities to detect and correct errors across six distinct scenarios beyond just comparing final accuracy before and after refinement. Furthermore, the benchmark analyzes the refinement performance by categorizing errors into six error types. Experiments with various open and closed MLLMs reveal bottlenecks and factors impeding refinement performance, highlighting areas for improvement in effective reasoning enhancement. Our code and dataset are publicly available at <https://github.com/naver-ai/MMRefine>.

1 Introduction

Recent advances have endowed MLLMs with remarkable capabilities, enabling them to tackle complex challenges such as mathematical reasoning and multimodal understanding (Grattafiori et al., 2024; Liu et al., 2024; Chen et al., 2025).

Instead of concentrating solely on scaling model parameters during training, current research aims to strengthen inference-time reasoning. Techniques such as *Self-Refinement*, where models iteratively improve their outputs (Madaan et al., 2023; Li et al., 2024a; Kumar et al., 2025; Huang et al., 2024b; Zhang et al., 2024), and engaging multiple models or agents in debate to achieve consensus (Liang et al., 2024; Talebirad and Nadiri, 2023; Chen et al., 2024) have gained traction.

These methodologies heavily rely on the ability of MLLMs to evaluate and refine their responses. If such capability is not sufficiently secured, refinement might unintentionally impair performance,

*Most work was done during the internship at NAVER Cloud AI.

†Corresponding author

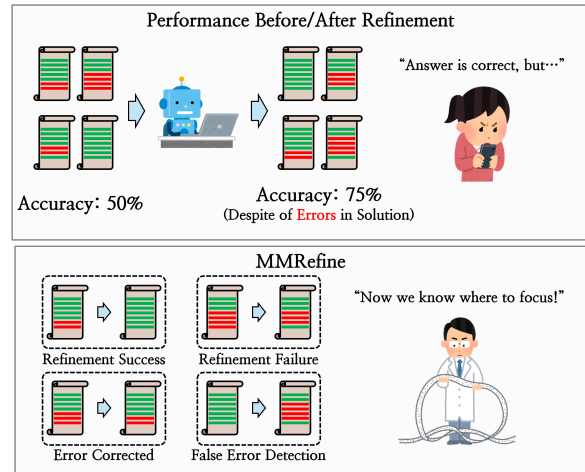


Figure 1: **Motivation.** Error refinement fails for a variety of scenarios and error types. Systematic evaluation is vital for providing accurate feedback and enhancing performance.

causing incorrect corrections and unnecessarily prolonged response times (Huang et al., 2024a). Therefore, it is essential to investigate whether MLLMs can accurately identify and correct errors in their reasoning processes.

However, previous studies primarily compare the accuracy of the final answer before and after applying refinement (Huang et al., 2024a; Li et al., 2024a) without sufficient analysis of MLLMs’ refinement capabilities. Although Yan et al. (2024) analyze specific abilities such as error localization and classification, their scope is limited to error detection capability.

In this paper, we propose a new benchmark, **MultiModal Refinement (MMRefine)**, to evaluate whether MLLMs can detect errors in their initial solutions and improve them. Unlike previous studies that compare accuracy before and after refinement, MMRefine assesses refinement outcomes beyond mere final accuracy by categorizing them into six scenarios that can occur during the refinement process: False Error Detection and Verification Suc-

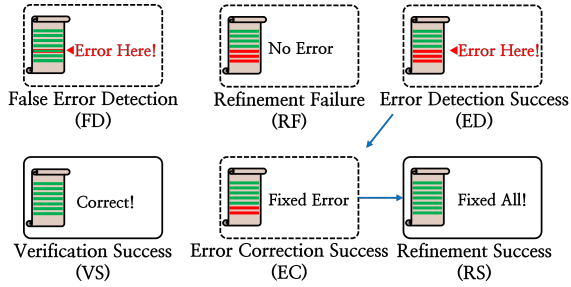


Figure 2: **Evaluation Protocol.** We define six scenarios to evaluate MLLM refinement capabilities.

cess for correct solutions, and Refinement Failure, Error Detection Success, Error Correction Success, and Refinement Success for incorrect solutions, as shown in Figure 1 and 2. Our approach enables the identification of refinement bottlenecks and offers a nuanced understanding of MLLMs’ refinement capabilities.

Through extensive experiments, we validate that MMRefine is effective for evaluating and analyzing the refinement capability of MLLMs. We evaluate 17 MLLMs’ refinement capability and examine which stages in the refinement process become bottlenecks for them. By comparing these scores with the actual self-reflection results on other benchmarks, we demonstrate the potential of the MMRefine as a reliable benchmark for refinement ability. Furthermore, we categorize the errors in MMRefine into six types and provide an analysis of the refinement performance for each error type. The analysis shows that MLLMs of various sizes and architectures exhibit varying strengths and weaknesses in correcting different types of errors.

Our study provides two main contributions. First, we introduce MMRefine, a MultiModal Refinement benchmark designed to systematically evaluate the refinement capabilities of MLLMs across the entire refinement process. Second, through comprehensive experiments, we evaluate the performance of each refinement process in MLLMs and analyze the error types to which they are vulnerable.

2 MMRefine

2.1 Overview of MMRefine

We propose MMRefine, a novel benchmark that is meticulously designed to evaluate the refinement capability of MLLMs. To effectively evaluate the models’ ability to refine their responses, we focus on mathematical problems that are sufficiently challenging to require refinement, often leading to

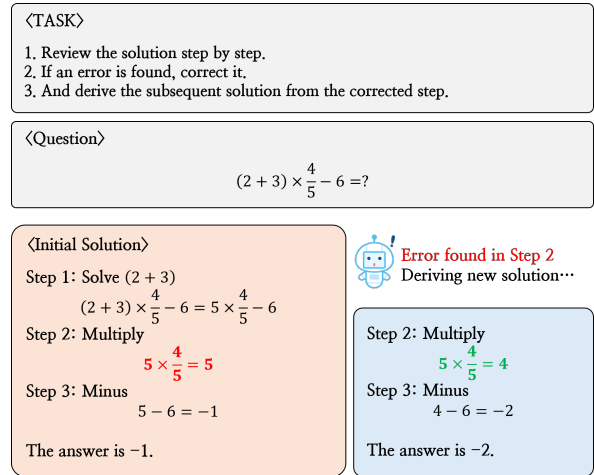


Figure 3: **Illustration of the Refinement Process.** The model identifies an **error** in the initial solution and **corrects** it, then proceeds to derive a revised solution from the point of correction.

longer reasoning paths, and allow for a clear and logical determination of correctness. Given an initial solution to a problem, we prompt the model to generate an improved solution as shown in Figure 3. Unlike previous studies (Huang et al., 2024a) that solely compared performance before and after refinement, we conduct a more granular evaluation by categorizing the refinement outcomes into six scenarios as depicted in Figure 2.

More specifically, we instruct MLLMs to review the solution step-by-step, identify and correct any detected errors, and regenerate the solution from the corrected point. Then, we categorize the refinement outcome into one of six scenarios by employing GPT-4O as a judge (Zheng et al., 2023) and considering the reference feedback. To ensure the reliability of LLM-based evaluations, we perform human verification and OPENAI O1 verification as described in Section 3.4. If the model incorrectly identifies an error in a correct solution, it is classified as **False Error Detection (FD)**; otherwise, it is categorized as **Verification Success (VS)**. If the model fails to detect an error in an incorrect solution, it is classified as **Refinement Failure (RF)**; otherwise, it is categorized as **Error Detection Success (ED)**. Note that since errors propagate to subsequent steps, we focus on the first error. Among the ED cases, if the error is accurately corrected, it is classified as **Error Correction Success (EC)**; furthermore, if the subsequent solution is generated flawlessly, it is categorized as **Refinement Success (RS)**. For details on the refinement and evaluation process, please refer to Appendix A.

	Refinement Failure (RF)	Error Detection Success (ED)	Error Correction Success (EC)	Refinement Success (RS)	False Error Detection (FD)	Verification Success (VS)	RefScore	mRecall
<i>Closed-Source MLLMs</i>								
GPT-4o (OpenAI et al., 2024a)	15.57	84.43	43.15	29.27	6.74	93.26	22.53	88.84
GEMINI-1.5-PRO (Google, 2024)	3.75	96.25	64.54	45.22	22.10	77.90	23.12	87.08
CLAUDE-3.5-SONNET (Anthropic, 2024)	27.95	72.05	32.65	18.95	6.74	93.26	12.21	82.65
<i>Open-Source MLLMs</i>								
LLAVA-ONEVISION-0.5B (Li et al., 2025)	36.40	63.60	2.06	2.06	75.66	24.34	-73.59	43.97
INTERNVL2.5-1B (Chen et al., 2025)	41.09	58.91	3.75	1.88	19.85	80.15	-17.97	69.53
QWEN2-VL-2B (Wang et al., 2024b)	51.59	48.41	3.19	2.44	19.10	80.90	-16.66	64.65
INTERNVL2.5-4B (Chen et al., 2025)	45.22	54.78	6.00	4.13	0.75	99.25	3.38	77.02
LLAVA-NEXT-7B (Liu et al., 2024)	42.40	57.60	5.44	4.13	4.49	95.51	-0.37	76.55
LLAVA-ONEVISION-7B (Li et al., 2025)	42.59	57.41	5.44	4.50	1.87	98.13	2.63	77.77
QWEN2-VL-7B (Wang et al., 2024b)	19.70	80.30	22.51	21.39	32.21	67.79	-10.82	74.05
INTERNVL2.5-8B (Chen et al., 2025)	25.14	74.86	11.44	5.82	10.49	89.51	-4.67	82.19
LLAMA-3.2-VISION-11B (Grattafiori et al., 2024)	22.14	77.86	16.14	10.51	32.96	67.04	-22.45	72.45
QWEN2-VL-72B (Wang et al., 2024b)	20.26	79.74	22.89	13.70	20.60	79.40	-6.90	79.57
LLAVA-NEXT-72B (Liu et al., 2024)	22.14	77.86	17.64	8.44	21.35	78.65	-12.91	78.26
LLAVA-ONEVISION-72B (Li et al., 2025)	31.14	68.86	21.76	11.07	4.87	95.13	6.20	81.99
INTERNVL2.5-78B (Chen et al., 2025)	15.57	84.43	32.65	20.26	17.98	82.02	2.29	83.23
LLAMA-3.2-VISION-90B (Grattafiori et al., 2024)	16.89	83.11	28.33	16.51	17.23	82.77	-0.72	82.94

Table 1: **MMRefine Benchmark Results.** The table shows performance metrics for closed-source and open-source MLLMs, with top scores highlighted in blue (closed-source) and green (open-source). Lower values are better for RF and FD, while higher values are better otherwise.

	Inference Time (s)		RefScore	Refinement Efficiency
	CoT	Refinement		
GPT-4o	11.22	7.77	22.5	0.33
GEMINI-1.5-PRO	7.67	7.78	23.1	0.23
CLAUDE-3.5-SONNET	6.06	4.82	12.2	0.15
LLAMA-3.2-VISION-11B	35.64	28.76	-22.5	-

Table 2: **MMRefine Refinement Efficiency.** We calculate refinement efficiency by dividing RefScore by the percentage of refinement inference time relative to the initial CoT inference time.

2.2 Evaluation Metrics

To analyze the bottleneck stages during the refinement process, we calculate the proportions for each result scenario. Since the ratio of incorrect solutions to correct solutions differs, we separately measure the ratios of FD and VS among the correct solutions and the ratios of RF, ED, EC, and RS among the incorrect solutions.

For straightforward comparison and evaluation of refinement capabilities, we introduce **RefScore**, a metric concentrating on the overall refinement performance of MLLMs, and **mRecall**, a metric emphasizing error detection performance. The RefScore is defined as:

$$\text{RefScore} = \text{RS} - \text{FD}$$

where RS and FD represent the proportions of corrected and uncorrected solutions, respectively. Meanwhile, mRecall is defined as:

$$\text{mRecall} = (\text{ED} + \text{VS})/2.$$

This measures the model’s ability to both detect actual errors and verify correct solutions accurately.

2.3 Dataset Construction

We construct MMRefine by carefully curating both text-only and visual math problems. We sample 100 text-only problems from MathOdyssey (Fang et al., 2024) and 100 visual problems from MathVision (Wang et al., 2024a) covering various subjects and levels of difficulty as described in Appendix B.

To provide a variety of initial solutions, we generate total 800 initial solutions using four MLLMs: GPT-4o (OpenAI et al., 2024a), GEMINI-1.5-PRO (Google, 2024; Team et al., 2024), CLAUDE-3.5-SONNET (Anthropic, 2024), and LLAMA-3.2-VISION-11B (Grattafiori et al., 2024). Note that MMRefine evaluates refinement processes under realistic conditions. Unlike previous studies that generated initial solutions by adding errors to correct solutions (Nath et al., 2025; Li et al., 2024b) or imposing constraints such as limiting the chain-of-thought steps on LLMs (Wu et al., 2024) to evaluate refinement capabilities, we employ solutions generated without any constraints.

For reliable evaluation, reference feedbacks are generated by OPENAI o1 (OpenAI et al., 2024b) using the original human-annotated solutions, and we validate them through the revision process. We test whether three MLLMs (GPT-4o, GEMINI-1.5-PRO, and CLAUDE-3.5-SONNET) can revise incorrect initial solutions when reference feedbacks are provided. We retain only the feedback where refinement success is confirmed across all models and ensure validity by either regenerating or manually correcting the flawed feedback.

To conduct a detailed analysis of MLLMs’ refine-

	MMRefine		MMRefine _{Text-only}		MATH-500		MMRefine _{Visual}		MathVista	
	RefScore	mRecall	RefScore	mRecall	CoT	Self-Reflection(Δ)	RefScore	mRecall	CoT	Self-Reflection(Δ)
GPT-4o	22.5	88.8	33.8	93.8	73.4	75.2 (+1.8)	12.9	84.5	60.5	61.2 (+0.7)
GEMINI-1.5-PRO	23.1	87.1	45.1	92.5	79.8	80.6 (+0.8)	-8.8	74.5	71.6	70.6 (-1.0)
CLAUDE-3.5-SONNET	12.2	82.7	21.3	88.3	61.2	62.2 (+1.0)	3.9	77.8	63.0	63.2 (+0.2)
LLAMA-3.2-VISION-11B	-22.5	72.5	-16.8	80.7	37.4	37.4 (0.0)	-13.7	73.6	48.4	47.3 (-1.1)

Table 3: **Comparison of MMRefine Scores and the Self-Reflection Results in MATH-500 and MathVista.** To conduct an in-depth analysis of the results in text-only and visual math problems, we report the results for the two subsets of MMRefine: MMRefine_{Text-only} consisting of MathOdyssey problems and MMRefine_{Visual} consisting of MathVision problems. Refer to Appendix E for details.

	Problem Understanding	Logical Reasoning	Calculation	Equation	Visual Perception	Spatial Reasoning
	<i>Closed-Source MLLMs</i>					
GPT-4o (OpenAI et al., 2024a)	36.7	29.4	32.8	34.7	26.3	11.5
GEMINI-1.5-PRO (Google, 2024)	36.7	48.6	67.2	61.2	35.0	23.1
CLAUDE-3.5-SONNET (Anthropic, 2024)	25.0	22.0	18.0	28.6	13.8	0.0
<i>Open-Source MLLMs</i>						
LLAVA-ONEVISION-0.5B (Li et al., 2025)	3.3	0.6	4.9	0.0	2.5	3.8
INTERNVL2.5-1B (Chen et al., 2025)	1.7	1.7	1.6	0.0	1.9	7.7
QWEN2-VL-2B (Wang et al., 2024b)	0.0	2.8	3.3	0.0	2.5	7.7
INTERNVL2.5-4B (Chen et al., 2025)	3.3	4.0	6.6	6.1	3.1	3.8
LLAVA-NEXT-7B (Liu et al., 2024)	5.0	1.7	3.3	2.0	3.8	26.9
LLAVA-ONEVISION-7B (Li et al., 2025)	3.3	2.8	4.9	4.1	4.4	19.2
QWEN2-VL-7B (Wang et al., 2024b)	11.7	19.8	26.2	10.2	26.3	34.6
INTERNVL2.5-8B (Chen et al., 2025)	5.0	4.5	8.2	8.2	6.3	0.0
LLAMA-3.2-VISION-11B (Grattafiori et al., 2024)	6.7	15.8	18.0	16.3	2.5	3.8
QWEN2-VL-72B (Wang et al., 2024b)	8.3	11.9	21.3	12.2	15.6	11.5
LLAVA-NEXT-72B (Liu et al., 2024)	8.3	7.3	9.8	12.2	8.1	7.7
LLAVA-ONEVISION-72B (Li et al., 2025)	15.0	11.9	14.8	8.2	8.8	7.7
INTERNVL2.5-78B (Chen et al., 2025)	16.7	20.9	26.2	16.3	21.3	11.5
Llama-3.2-Vision-90B (Grattafiori et al., 2024)	15.0	16.4	19.7	18.4	15.6	15.4

Table 4: **Comparison of RefScore by First Error Type.** Maximum and minimum values are highlighted.

ment capabilities based on error types, we manually categorize the first errors in the initial solutions into six types as depicted in Appendix C. Detailed explanations of six error types are detailed in Appendix D.

3 Experiments

3.1 Overall Performance

We evaluate 17 MLLMs, including 3 closed-source models and various open-source models from 5 MLLM families of differing sizes, as shown in Table 1. In terms of mRecall, closed-source models demonstrate superior performance compared to all open-source models below 11B. Only a few large-scale open-source models, namely INTERNVL2.5-78B and LLAMA-3.2-VISION-90B manage to surpass CLAUDE-3.5-SONNET. Additionally, closed-source models correct errors in over 32% of initial solutions, while most open-source models correct less than 20%. For RS, which reflects perfect response improvement, only two open-source models, INTERNVL2.5-78B and QWEN2-VL-7B, exceed the performance of the lowest closed-source

records. Despite this, their RefScores remain lower than closed-source models due to high FD. These findings suggest that the current error correction proficiency of most open-source models remains inadequate for effective refinement, even in larger models exceeding 70B parameters. However, notable exceptions including QWEN2-VL-7B, which achieves a higher RS score than CLAUDE-3.5-SONNET, and INTERNVL2.5-8B, which records a high mRecall score of 82.19, indicate refinement potentials even in mid-scale models.

To ascertain which MLLMs offer reasonable refinement performance relative to the increased computational cost caused by refinement, we also measure the refinement efficiency, as shown in Table 2. Adding the refinement step increases inference time by 60-100% compared to the initial CoT inference. Notably, refinement efficiency differs significantly between models. Although GEMINI-1.5-PRO achieves a higher RefScore compared to GPT-4o, the refinement efficiency of GPT-4o is substantially higher. In practice, adopting refinement necessitates balancing the increase in inference time with the anticipated performance gain.

3.2 Correlation with Self-Reflection

We also analyze the correlation between MMRefine scores and score changes after self-reflection in other math benchmarks: MATH-500 (Hendrycks et al., 2021) and MathVista (Lu et al., 2024), as shown in Table 3. The results show that RefScore and mRecall are correlated with the models’ refinement capability. With the exception of GEMINI-1.5-PRO, RefScores in text-only and visual math problems are directly correlated with the score changes in MATH-500 and MathVista (correlation coefficient 0.82). GEMINI-1.5-PRO, particularly for visual problems, exhibits a relatively low mRecall, which appears to have led to a decrease in scores after self-reflection in MathVista. From the results, we demonstrate that the MMRefine scores are valuable indicators of the refinement capability.

3.3 Error Type Analysis

To understand what types of errors impede effective refinement, we analyze the RefScore by six distinct error types, as detailed in Table 4. While different MLLMs exhibit varying strengths and weaknesses in refining specific error types, our findings indicate that larger models with higher capacities tend to perform significantly better at correcting four text-related errors than two image-related ones. In contrast, models smaller than 7B often demonstrate superior handling of image-related errors. For instance, LLAVA-NEXT-7B and QWEN2-VL-7B perform better than even closed-source models in correcting spatial reasoning errors. While this discrepancy may be partially attributed to differences in LLM and vision encoder sizes, it could also be influenced by the curriculum through which MLLMs acquire their capabilities.

We also compare the correlation between RefScores for each error type. As shown in Figure 4, the performance on most error types is highly correlated with that on other error types, whereas spatial reasoning error type shows low correlations overall. This observation suggests that there may be alternative approaches to enhancing refinement performance for specific error types, such as spatial reasoning errors, beyond merely scaling up the refinement capability of MLLMs.

3.4 LLM-based Evaluation

The nature of solving math problems allows for diverse approaches and infinitely varied errors and correction methods within the reasoning process.

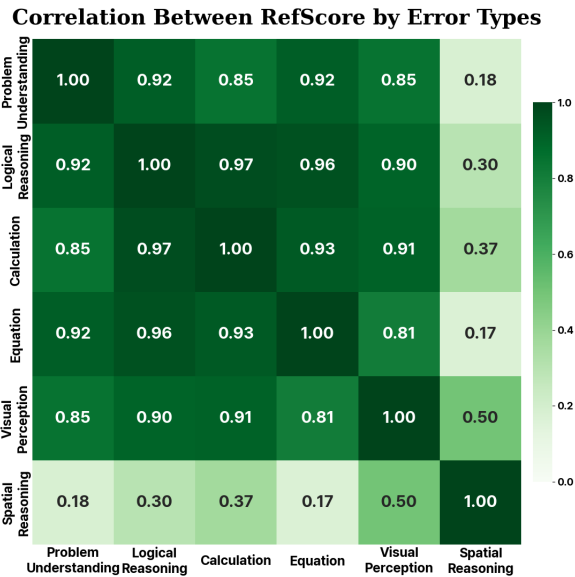


Figure 4: **Correlation Between Refscore by Error Types.** We calculate the correlation coefficients of RefScore for each error type across 17 models.

Because the human evaluation of such complex reasoning is not only highly demanding but can even be inaccurate, automated methods such as LLM-as-a-Judge (Zheng et al., 2023) have been proposed to evaluate MLLMs’ reasoning processes. In MMRefine, we use GPT-4O as a judge. To ensure the reliability of it, we compare GPT-4O’s judgments with those from human evaluations and OPENAI O1. GPT-4O achieves 72% agreement with human judgments and 73% agreement with the OPENAI O1 judgments. While the alignments are not perfect, the practical advantage of LLM-based evaluation becomes clear when considering its efficiency and scalability. Whereas a human evaluation conducted by an expert with university-level mathematical knowledge takes over 8 hours, GPT-4O provides reliable judgments in much shorter time and effort.

4 Conclusion

This paper introduces MMRefine, which evaluates the refinement capabilities of MLLMs through an analysis of their outcomes across six distinct scenarios and six error types. Our comprehensive assessment of 17 MLLMs reveals that larger models tend to refine textual errors better, whereas smaller models are more effective with visual errors. For spatial reasoning errors, only specific models exhibit a certain level of refinement capability. These insights into intrinsic refinement capabilities can enhance MLLMs’ reasoning abilities and provide guidance for addressing their weaknesses.

Limitations

In this paper, we generate evaluation data by solving problems collected from two math benchmark datasets using four closed-source and open-source MLLMs. While this approach allows for consistent evaluation of MLLM refinement capability, the resulting data inherently lacks the diversity of real-world use case scenarios and the breadth of initial solutions common in practice, such as solutions from other models, human-authored responses, and non-mathematical reasoning processes. Furthermore, although various correct answers may exist when solving math problems, we adopt a single reference solution and conduct rigorous evaluations to facilitate LLM-based assessment. This bias can lead the model to be robust only in a few mathematical reasoning methods and overlook other important issues (e.g., fact verification, diverse and original problem-solving approaches). Although we inevitably conduct evaluations based on a limited reference solution to enhance the credibility of LLM-as-judge assessment, we aim to explore more flexible methods for evaluating and verifying the reasoning validity of MLLMs in future work.

Ethical Considerations

We acknowledge that, due to practical considerations, the experimental results detailed in this paper are derived from single-run assessments. However, to uphold the reliability of our evaluation framework, we dedicate considerable effort to refine the prompts for the LLM-based judges carefully. Additionally, for the selected model, most prominently GEMINI-1.5-PRO, we execute evaluations across three iterations and observe that the standard deviation of the resulting RefScore remains comfortably below 1, thus suggesting a degree of score stability.

References

Anthropic. 2024. [Claude 3.5 sonnet model card addendum](#).

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024. [Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning](#). *Preprint*, arXiv:2409.12147.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi

Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. [Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data](#). *Preprint*, arXiv:2406.18321.

Google. 2024. [Gemini 1.5 pro](#).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimppoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal,

Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gouget, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiyen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie DelPierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weiss-

man, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabza, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangrabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihalescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiao Cheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).

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](#). In *The Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. [Large language models cannot self-correct reasoning yet](#). In *The Twelfth International Conference on Learning Representations*.
- Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. 2024b. [QueryAgent: A reliable and efficient reasoning framework with environmental feedback based self-correction](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5014–5035, Bangkok, Thailand. Association for Computational Linguistics.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2025. [Training language models to self-correct via reinforcement learning](#). In *The Thirteenth International Conference on Learning Representations*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [Llava-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024a. [Confidence matters: Revisiting intrinsic self-correction capabilities of large language models](#). *Preprint*, arXiv:2402.12563.
- Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024b. [Evaluating mathematical reasoning of large language models: A focus on error identification and correction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11316–11360, Bangkok, Thailand. Association for Computational Linguistics.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, 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](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, and Mitesh M. Khapra. 2025. [Can vision-language models evaluate handwritten math?](#) *Preprint*, arXiv:2501.07244.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew Mishchenko, Angela Baek, Angela Jiang, Antoine Peltisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kelloog, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan

Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lillian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Fevrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeih, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Ra-

jan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shiron Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. *Gpt-4o system card*. *Preprint*, arXiv:2410.21276.

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, Vitthyr 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. 2024b. [Openai ol system card](#). *Preprint*, arXiv:2412.16720.

Yashar Talebirad and Amirhossein Nadiri. 2023. [Multi-agent collaboration: Harnessing the power of intelligent llm agents](#). *Preprint*, arXiv:2306.03314.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Serincinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Hari-

dasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Henigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Or Gil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurusurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillcrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa,

Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Challenge Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkels-son, Marcello Maggioni, Daniel Zheng, Yury Sul-sky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjoes, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlias, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Cave-ness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Kar-mar, Lev Proleev, Abe Ittycheriah, Soheil Has-sas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin John-son, Behnam Neyshabur, Justin Mao-Jones, Ren-shen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, So-

phie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Se-bastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangoeei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Ku-mar, Colton Bishop, Adams Yu, Sarah Hodgkin-son, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Char-lotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Laksh-minarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Re-casens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Erbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Pol-acek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shan-tanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Pa-tel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jes-sica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chau-mont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bu-lanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Gar-rette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Itur-rate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Va-

hab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeewan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Iliia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Ram-mohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis,

Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnappalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkupati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srinu Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Viji, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemny, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan

- Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petriani, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Meray, Martin Baeuml, Trevor Strohmaier, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. [Measuring multimodal mathematical reasoning with MATH-vision dataset](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. 2025. [Visualprm: An effective process reward model for multimodal reasoning](#). *Preprint*, arXiv:2503.10291.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [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.
- Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. 2024. [Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning](#). *Preprint*, arXiv:2412.02172.
- Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, Aoxiao Zhong, Kun Wang, Hui Xiong, Philip S. Yu, Xuming Hu, and Qingsong Wen. 2024. [Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection](#). *Preprint*, arXiv:2410.04509.
- Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. [Small language models need strong verifiers to self-correct reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15637–15653, Bangkok, Thailand. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Refinement & Evaluation Protocol Details

In this section, we delve into the specifics of how refinement outcomes are generated and evaluated using the MMRefine benchmark. Initially, we provide the model with a math problem and an initial solution and instruct it to review the solution step-by-step using the prompt in Figure 5. If an error is identified during the review, the model stops reviewing and performs refinement starting from that step. At the end of refinement, the model provides the following outputs: the correctness of the initial solution, the explanation for this determination, and the final answer.

When the initial solution refined by the model is indeed correct, we determine whether the refinement result is Verification Success (VS) or False Error Detection (FD) by parsing the correctness from the model’s response, as shown in Figure 6. Conversely, if the initial solution is incorrect, we evaluate the model’s refinement outcome by comparing it with the reference feedback through the prompt in Figure 7. Specifically, the model’s success begins with Error Detection (ED), which is a prerequisite for subsequent scenarios, assessed based on the “Error Detection” rubric of the prompt. Following successful error detection, the model’s Error Correction (EC) capability is evaluated using the “Error Correction” rubric. Finally, Refinement Success (RS), determined based on the “Effectiveness and Correctness of the Feedback” rubric, signifies that the model not only detects and corrects the error(s) but also derives a correct solution to the correct answer, encompassing both ED and EC.

B Datasets Details

We sample 200 mathematical questions from the MathOdyssey (Fang et al., 2024) and MathVision (Wang et al., 2024a) dataset as described in Section 2.3. The MathOdyssey dataset features mathematical questions from a wide range of subjects, encompassing difficulty levels from High School to University and Olympiad. Conversely, the MathVision dataset offers visual math problems across various domains, categorized by difficulty levels 1 through 5. We curate 100 problems from each of these two datasets to construct our benchmark, as summarized in Table 5.

MathOdyssey dataset is distributed under the CC BY-SA 4.0 license, which permits its use as a test set. The license covers the dataset itself but

	Subject	# of Questions
MathOdyssey	Algebra	39
	Precalculus	12
	Geometry	11
	Combinatorics	10
	Linear Algebra And Abstract Algebra	7
	Calculus And Analysis	6
	Probability	5
	Differential Equations	4
	Statistics	4
	Number Theory	1
	Calculus	1
	Level	# of Problems
	High School Math	35
	High School Competition	39
	College Math	26
	Subject	# of Questions
MathVision	Metric Geometry	48
	Solid Geometry	13
	Combinatorial Geometry	7
	Algebra	6
	Transformation Geometry	6
	Descriptive Geometry	6
	Combinatorics	5
	Graph Theory	3
	Logic	3
	Arithmetic	2
	Counting	1
	Level	# of Problems
	Level 2	25
	Level 3	30
	Level 4	29
	Level 5	16

Table 5: **MMRefine Data Statistics.** MMRefine consists of problems that cover a wide range of subjects and levels of difficulty.

not questions in the dataset. The MathVista dataset is available under the MIT License. MMRefine, derived from two benchmarks, is released under the CC BY-SA 4.0 license. This license covers the elements we create or label, while the copyright of the original questions remains with their respective authors. Similar to MathOdyssey, MMRefine is also restricted for testing purposes only, and its use as training data for models is prohibited.

C Human Annotations Details

As detailed in Section 2.3, we manually annotate the first error type in each initial solution. Annotators are tasked with labeling each initial solution, referencing the math problem, the model’s generated solution, and the problem’s original solution to determine the presence of errors and, if errors are found, to categorize the first error by its type. The annotation is conducted by 14 annotators, with 12 holding a bachelor’s degree and 2 holding a master’s degree.

	Source of Initial Solution			
	GPT-4O	GEMINI-1.5-PRO	CLAUDE-3.5-SONNET	LLAMA-3.2-VISION
GPT-4O	20.97	14.82	16.90	33.41
GEMINI-1.5-PRO	18.25	11.86	21.79	38.06
CLAUDE-3.5-SONNET	19.65	7.06	3.69	18.49
LLAMA-3.2-VISION-11B	-15.83	-24.51	-30.01	-19.50

Table 6: **Performance Comparison of MLLMs Across Different Models Generating Initial Solutions.** In each row, the highest RefScore is highlighted in blue, the second highest in green, and the third highest in pink.

D Detailed Explanation of Error Types

To enable a nuanced analysis of MLLMs’ refinement capabilities across various situations, particularly concerning the nature of errors, we implement a categorization scheme encompassing six distinct error types. **Problem understanding error** occurs when the model misinterprets the instructions or constraints explicitly stated in the problem description. **Logical reasoning error** denotes instances where the solution exhibits a flaw in the logical flow of argumentation, leading to an invalid conclusion. **Calculation error** refers to inaccuracies arising from numerical computation mistakes within the mathematical derivation. **Equation error** encompasses a range of mistakes related to algebraic manipulation, including, but not limited to, incorrect equation expansion or invalid variable substitution. **Visual perception error** is identified when the model fails to correctly interpret or recognize essential information conveyed through the problem’s accompanying image. Lastly, **spatial reasoning error** is characterized by errors stemming from flawed spatial reasoning, such as incorrect assessments of geometric relationships or misinterpretations of spatial limitations. The error type distribution of MMRefine is presented in Figure 8 and Figure 9.

E Experimental Details for Section 3.2

To explore the correlation of MMRefine with existing benchmarks, we conduct self-reflection experiments on MATH (Hendrycks et al., 2021) and MathVista (Lu et al., 2024), prominent benchmarks within the Large Language Model research community. For MATH, we perform evaluations using the 500 test subset, as used in (Lightman et al., 2024). For MathVista, evaluations are conducted on the testmini set. We begin by evaluating the model’s baseline performance using basic Chain-of-Thought (CoT) prompting (Wei et al., 2022).

Subsequently, we prompt the model to refine its initial response through self-reflection, utilizing the prompt detailed in Figure 10.

F RefScore Comparison by Solution Source

We conduct experiments to investigate how refinement efficacy varies depending on the model that provides the initial solution. As shown in Table 6, all models achieve their best RefScore from initial solutions originating from LLAMA-3.2-VISION or GPT-4O. Interestingly, most models tend to successfully refine initial solutions generated by LLAMA-3.2-VISION. A plausible interpretation for this trend is that LLAMA-3.2-VISION tends to generate responses with errors skewed towards easier problem instances, thereby facilitating more effective refinement, as shown in Figure 11.

G Comparison with Process Reward Models

We evaluate whether RefScore correlates with Process Reward Models (PRMs), which are used to assess MLLMs’ reasoning processes and select better ones. As shown in Table 8, we calculate the correlation between RefScore and the directional changes in rewards of VISUALPRM-8B (Wang et al., 2025) before and after refinement. Our experimental results show a moderate relationship between assessing improvements from refinements using PRM rewards and the RefScore. This finding can shed light on new directions for future research to analyze and enhance the performance of reward models in selecting better responses, particularly from the refinement perspective.

We also examine whether PRMs could also improve the refinement process, as shown in Table 7. After applying the best-of-N selection with the VISUALPRM-8B, we observe a trade-off where Error Detection (ED) decreases while Verification

	RF (↓)	ED (↑)	EC (↑)	RS (↑)	VS (↑)	FD (↓)	RefScore	mRecall
GPT-4O	15.57	84.43	43.15	29.27	93.26	6.74	22.53	88.84
+VISUALPRM-8B (N=4)	20.83	79.17	40.53	22.33	96.63	3.37	18.96	87.90
GEMINI-1.5-PRO	3.75	96.25	64.54	45.22	77.90	22.10	23.12	87.08
+VISUALPRM-8B (N=4)	7.88	92.12	58.91	39.59	82.40	17.60	21.98	87.26
CLAUDE-3.5-SONNET	27.95	72.05	32.65	18.95	93.26	6.74	12.21	82.65
+VISUALPRM-8B (N=4)	31.52	68.48	33.58	18.76	95.51	4.49	14.27	81.99
LLAMA-3.2-VISION-11B	22.14	77.86	16.14	10.51	67.04	32.96	-22.45	72.45
+VISUALPRM-8B (N=4)	25.52	74.48	21.95	13.32	80.15	19.85	-6.53	77.32

Table 7: **MMRefine Performance Before and After Applying Best-of-N Selection with the VisualPRM.** Improved values are highlighted in **bold**.

	RefScore	Reward Change	Threshold	RF (↓)	ED (↑)	VS (↑)	FD (↓)	mRecall
GPT-4O	22.5	0.14	0.1	71.11	28.89	61.42	38.58	45.16
GEMINI-1.5-PRO	23.1	-0.12	0.2	75.99	24.02	70.04	29.96	47.03
CLAUDE-3.5-SONNET	12.2	0.11	0.3	78.05	21.95	76.78	23.22	49.37
LLAMA-3.2-VISION-11B	-22.5	-0.10	0.4	81.99	18.01	81.65	18.35	49.83
Correlation Coefficient: 0.4292			0.5	84.43	15.57	86.14	13.86	50.86
			0.6	86.68	13.32	90.64	9.36	51.98
			0.7	91.37	8.63	94.38	5.62	51.51
			0.8	95.87	4.13	98.50	1.50	51.31
			0.9	99.62	0.38	99.63	0.37	50.00

Table 8: **Correlation Between RefScore and VisualPRM.** We measure the correlation between RefScore and the directional changes in VisualPRM rewards before and after refinement (+1 for increase, -1 for decrease, 0 for no change).

Table 9: **Error Detection Performance of VisualPRM.** We determine that a step is incorrect when the probability of “incorrect” exceeds that of “correct” by a certain threshold.

Success (VS) increases. This results in performance improvements for models with low VS but can lead to a decrease in performance for those with high VS due to reduced ED. These findings highlight the potential of PRMs to assist with the refinement of MLLMs that lack inherent error detection and correction abilities.

Furthermore, we conduct an experiment to measure the MMRefine performance of the PRM itself. Since the PRM only outputs a reward score for each step, we evaluate its error detection ability using the method described in the VISUALPRM paper, as shown in Table 9. The threshold ablation shows optimal mRecall (51.98) around a threshold of 0.6. Consistent with the findings in Table 7, the results indicate that the PRM exhibits high Verification Success (VS), suggesting strong robustness against false error detections but low Error Detection (ED).

ment Failure (RF). On the other hand, Figure 13 shows GEMINI-1.5-PRO trying to find and correct an error in the initial solution generated by CLAUDE-3.5-SONNET; however, it does not successfully rectify the error. Despite the final answer being A, which aligns with the ground truth, this failure leads MMRefine to classify this instance as an Error Detection Success (ED).

H Qualitative Examples

Figure 12 and 13 illustrate the outcomes of the models’ refinement attempts on the MMRefine benchmark. In Figure 12, CLAUDE-3.5-SONNET attempts to refine the initial solution from GPT-4O but fails to identify any errors, incorrectly judging it as a correct solution, which results in a Refine-

You are a mathematical expert with extensive knowledge across various mathematical fields. Your task is to meticulously evaluate and, if necessary, correct a given mathematical question and its proposed solution.

Follow these steps:

1. Carefully read the provided question and solution.
2. Conduct a step-by-step review of the solution, addressing the following for each step:
 - Verify the mathematical correctness and logical flow.
 - Identify any errors including calculation errors, misunderstanding of the problem, or reasoning error.
 - If an error is found, immediately stop the review process and proceed to step 3.
 - If no error is found, continue to the next step.
3. If an error is found:
 - Provide a brief explanation of the error.
 - Correct the solution starting from the erroneous step.
 - Complete the rest of the solution correctly.
4. If no errors are found in the entire solution, provide a brief confirmation of its correctness.

Output your analysis in the following format:

Review and Correction (if applicable):

Step 1: [Brief assessment of step 1]

Step 2: [Brief assessment of step 2]

...

Step X: [Brief assessment of step X]

Error found in step X: [Brief explanation of the error]

Corrected solution from step X:

Step X: [Corrected step]

Step X+1: [Next correct step]

...

[Final step]

...

Overall Assessment:

Correctness: [Correct / Incorrect]

Explanation: [Concise explanation of the assessment]

Final Answer: [Correct final answer]

Question: {question}

Solution: {initial_solution}

Figure 5: Prompt to Perform Refinement

Given the model's response, output 1 if 'Correctness' is 'correct', otherwise output 0. Respond with only the number.

Model's Response: {response}

Output:

Figure 6: Prompt to Evaluate Samples with Correct Initial Solutions

You are an expert evaluator assessing the quality of feedback provided on an initial solution to a problem. Your task is to determine if the feedback is effective in guiding the initial solution towards a correct answer. You will be provided with three components:

1. **Initial Solution:** The initial attempt at solving the problem.
2. **Feedback:** Specific feedback provided in response to the initial solution.
3. **Reference Feedback:** A verified, high-quality feedback to the initial solution.

Your evaluation should consider the following aspects:

- **Error Detection:** Does the feedback correctly identify the errors or shortcomings in the initial solution?
- **Error Correction:** Does the feedback effectively address the problems in the initial solution?
- **Effectiveness and Correctness of the Feedback:** Does the feedback guide the initial solution towards the correct answer efficiently? Does it reach the same answer and logic as the reference feedback in terms of its core principles?

Output your assessment in the following format:

Error Detection: [0/1]
Error Correction: [0/1]
Effectiveness and Correctness of the Feedback: [0/1]

No additional feedback or comment is required.

Initial Solution: {initial_solution}
Feedback: {feedback}
Reference Feedback: {reference_feedback}

Output:

Figure 7: Prompt to Evaluate Samples with Incorrect Initial Solutions

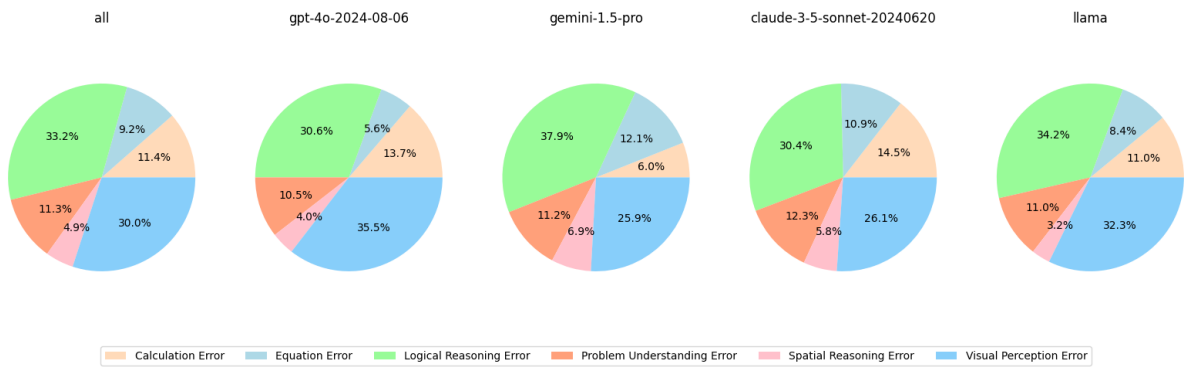


Figure 8: Error Type Distribution of Initial Solutions by Model

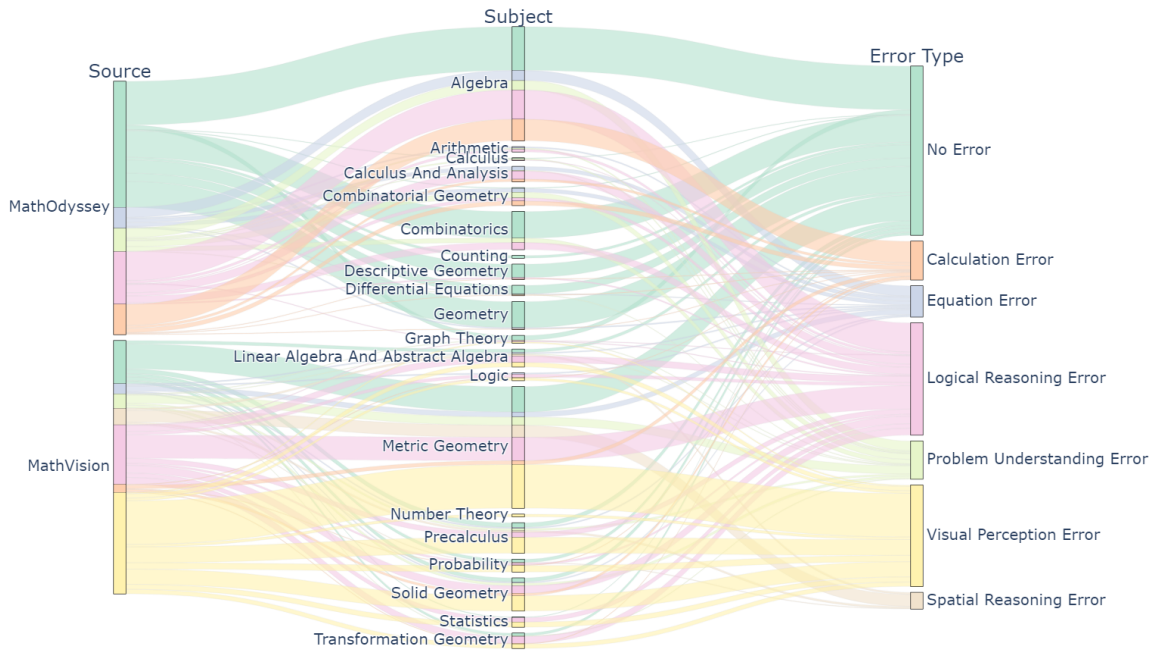


Figure 9: Problem Source, Subject, and Error Type Distribution in MMRefine

Review your previous reasoning about the question, then finally answer the question.

Question: {question}

Your Previous Solution: {previous_solution}

Figure 10: Prompt to Perform Self-Reflection

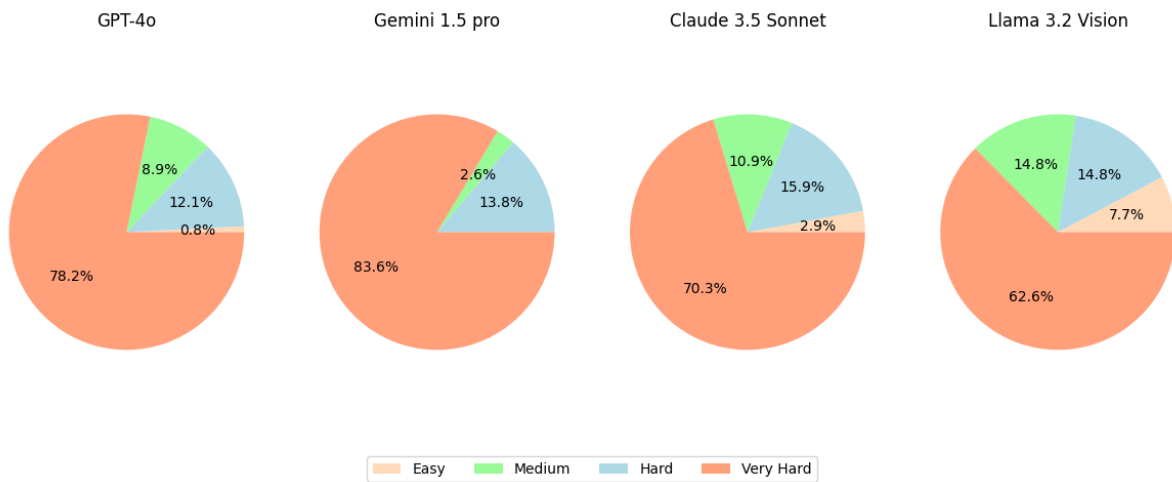
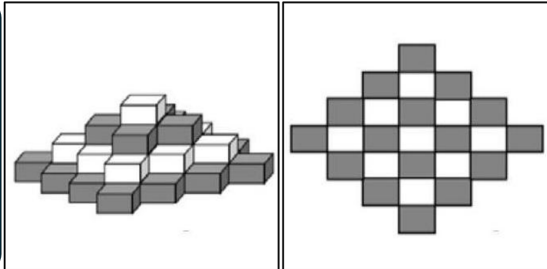


Figure 11: **Difficulty Distribution of Problems for Which Initial Solutions Are Incorrect by Model.** Problem difficulty is determined by the number of MLLMs that correctly solve it. Specifically, if three out of the four models (GPT-4o, GEMINI-1.5-PRO, CLAUDE-3.5-SONNET, and LLAMA-3.2-VISION) solve a problem correctly, the difficulty is categorized as 'Easy'. If two models solve it, the difficulty is 'Medium' and so on.

<QUESTION>

The two diagrams show a side view and a plan view of a tower made with light and dark coloured blocks. In the tower, only dark coloured blocks are placed on top of dark coloured blocks and only light coloured blocks are placed on top of light coloured blocks. How many blocks in the tower are light coloured?

Answer: 20



<INITIAL SOLUTION> (by GPT-4o)

To determine the number of light-colored blocks in the tower, let's analyze the diagrams step-by-step:

1. Plan View Analysis:

- The plan view (right diagram) shows a top-down view of the tower.
- Identify the pattern: light and dark blocks alternate.
- The central position is light-colored, and blocks alternate outward from there.

2. Side View Analysis:

- The side view (left diagram) provides a perspective on the layer arrangement and height.
- Count the layers from bottom to top, observing the alternating pattern.

3. Layer-by-Layer Analysis:

- From the plan view, we see a diamond shape pattern, alternating colors.
- Bottom Layer: Count the number of light-colored blocks.
- Alternating rows start and end with light-colored blocks.
- Middle Layers: Follow the alternating pattern, ensuring consistency with "light on light" and "dark on dark" placement.
- Top Layer: Usually the smallest number of blocks, following the pattern.

4. Count Each Layer:

- Bottom Layer: 9 blocks (5 light, 4 dark).
- Next Layer: 7 blocks (4 light, 3 dark).
- Next Layer: 5 blocks (3 light, 2 dark).
- Next Layer: 3 blocks (2 light, 1 dark).
- Top Layer: 1 block (1 light). (Error in Step 4)

5. Summing Light-Colored Blocks:

- Add the light-colored blocks from each layer.
- $(5 + 4 + 3 + 2 + 1 = 15)$.

Therefore, there are 15 light-colored blocks in the tower.

<Refinement> (by Claude 3.5 Sonnet)

I'll review the solution step-by-step:

Review and Correction:

Step 1: Correct analysis of the plan view.

Step 2: Correct analysis of the side view.

Step 3: Correct approach to layer-by-layer analysis.

Step 4: Correct count of blocks in each layer. (Error Detection Failed)

Step 5: Correct summation of light-colored blocks.

Overall Assessment:

Correctness: Correct

Explanation: The solution correctly analyzes both the plan view and side view of the tower. It accurately counts the number of light-colored blocks in each layer and sums them up correctly.

Final Answer: 15 light-colored blocks

The solution is methodical and accurate. It correctly interprets the alternating pattern of light and dark blocks, both horizontally and vertically. The count of light-colored blocks in each layer (5, 4, 3, 2, 1) is accurate, and their sum (15) is correct. The explanation also adheres to the rule that light-colored blocks are only placed on top of light-colored blocks, which is crucial for the correct solution.

Figure 12: Refinement Failure Example. Model (CLAUDE 3.5 SONNET) fails to detect an error in step 4 of initial solution.

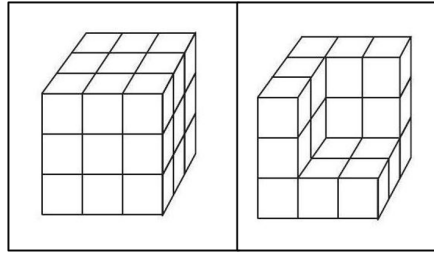
<QUESTION>

It takes 9 litres of paint to cover the surface of the cube on the left.

How much paint would it take to cover the surface of the shape on the right?

A. 9 litres, B. 8 litres, C. 6 litres, D. 4 litres, E. 2 litres

(Answer: A)



<INITIAL SOLUTION> (by Claude 3.5 Sonnet)

Let's approach this systematically:

- 1) First, let's consider the cube on the left:
 - It's a 3x3x3 cube
 - Its surface area consists of 6 faces, each with 9 small squares
 - Total surface area = $6 * 9 = 54$ small squares
 - 2) Now, let's look at the shape on the right:
 - It's made up of $3x3x3 = 27$ small cubes, but some are removed
 - We need to count the exposed faces
 - 3) Counting exposed faces on the right shape:
 - Top: 9 squares - Bottom: 9 squares - Front: 7 squares
 - Back: 7 squares - Left: 7 squares - Right: 7 squares
 - Total: $9 + 9 + 7 + 7 + 7 + 7 = 46$ squares
 - 4) Comparing the two:
 - Left cube: 54 squares need 9 litres - Right shape: 46 squares
 - 5) Calculating the paint needed:
 - If 54 squares need 9 litres, then 1 square needs $9/54 = 1/6$ litre
 - 46 squares will need: $46 * (1/6) = 46/6 = 7.67$ litres
 - 6) Rounding to the nearest option: 7.67 litres is closest to 8 litres
- Therefore, the answer is B: 8 litres.

<Refinement> (by Gemini 1.5 Pro)

Review and Correction (if applicable):

Step 1: Correct. The left cube is 3x3x3 and has a surface area of 54 small squares.

Step 2: Correct. The right shape is derived from a 3x3x3 cube with some cubes removed.

Step 3: **Incorrect. The count of exposed faces on the right shape is inaccurate.**

(Error Detection Success)

Error found in step 3: The solution incorrectly counts only the outer faces. The inner faces of the "carved out" section also need to be painted.

Corrected solution from step 3:

Step 3: Counting exposed faces on the right shape:

- Top: 9 squares - Bottom: 9 squares - Front: 7 squares
- Back: 7 squares - Left: 7 squares - Right: 7 squares
- Inner surfaces (top): 2 squares - Inner surfaces (left): 2 squares - Inner surfaces (back): 2 squares
- Total: $9 + 9 + 7 + 7 + 7 + 7 + 2 + 2 + 2 = 52$ squares

(Error correction failed due to visual perception error)

Step 4: Comparing the two:

- Left cube: 54 squares need 9 litres - Right shape: 52 squares

Step 5: Calculating the paint needed:

- If 54 squares need 9 litres, then 1 square needs $9/54 = 1/6$ litre
- 52 squares will need: $52 * (1/6) = 52/6 = 26/3 = 8.67$ litres (approximately)

Step 6: Rounding to the nearest option:

- 8.67 litres is closest to 9 litres.

Overall Assessment:

Correctness: Incorrect

Explanation: The original solution failed to account for the inner surfaces created by the removal of the smaller cubes within the larger cube. This resulted in an underestimation of the total surface area to be painted and, consequently, an incorrect calculation of the required paint.

Final Answer: 9 litres (A)

(Refinement failed, despite a correct answer, because of errors in the solution process.)

Figure 13: **Error Detection Success Example.** Model (GEMINI-1.5-PRO) manages to detect the initial error but fails to correct it due to a visual perception error in the refinement process.