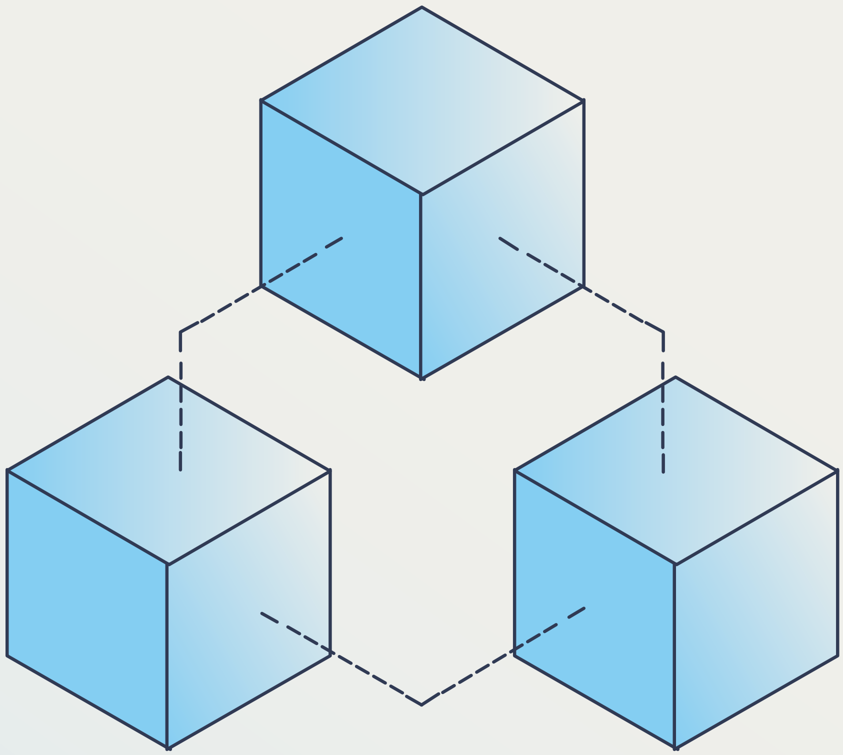
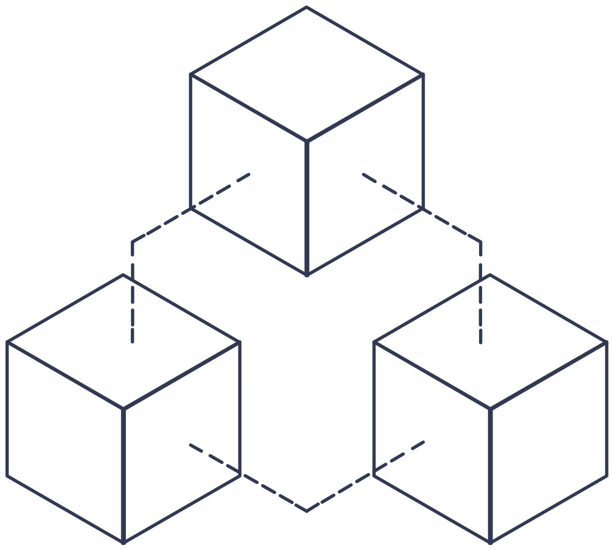


Leveraging Chain-of-Thought to Improve LLM Reasoning – A Case Study on Mathematical Reasoning Data

Guide to CoT reasoning for LLMs featuring an expert case study on how Appen built a mathematical reasoning dataset for a leading technology company.

Appen



Leveraging Chain-of-Thought to Improve LLM Reasoning – A Case Study on Mathematical Reasoning Data

Guide to CoT reasoning for LLMs featuring an expert case study on how Appen built a mathematical reasoning dataset for a leading technology company.

Appen

Copyright © Appen 2024 – All Rights Reserved

Table of Contents

LLM Reasoning: A Challenge of Modern AI

Large Language Models (LLMs) have taken the world by storm with their remarkable ability to understand and generate human-like text across an incredible breadth of domains. These powerful AI systems, trained on vast datasets of online data, can perform a wide range of tasks including answering questions, summarizing documents, writing poems, and even writing code.

However, one area that remains an outstanding challenge for large language models is robust reasoning - the ability to take a set of facts or premises, logically combine them, and arrive at valid conclusions. Reasoning is a critical capability not just for question-answering, but for any task that involves taking information and using it to make judicious decisions, develop well-supported arguments, or even output plans for solving complex tasks by breaking them down into easier sub-tasks.

Evaluating the consistency and reasoning abilities of LLMs reveals that while proprietary models generally outperform public ones, none of them consistently achieve high scores in both consistency and reasoning (source: Saxena et al. 2024). Thus, understanding and improving reasoning in LLMs remain critical for their continued development and reliable use.

In this guide, we'll explore what makes reasoning such a difficult frontier for modern AI and explore how to build Chain-of-Thought (CoT) reasoning data using a case study on mathematical reasoning data produced by our team of experts.

How chain-of-thought prompting elicits reasoning in large language models

Just as school children are instructed to “show their work” in math class, so too can LLMs benefit from clearly demonstrating the logical thought processes behind their outputs. With chain-of-thought prompting, LLM reasoning is made clear as the model articulates the logic behind the final output step-by-step.

Compared to standard prompting where a unit consists of <Question, Answer>, chain-of-thought prompting consists of <Question, CoT, Answer> triples (see Figure 1). This systematic approach enhances the LLM’s reasoning and accuracy by capturing the logical progression of thoughts.

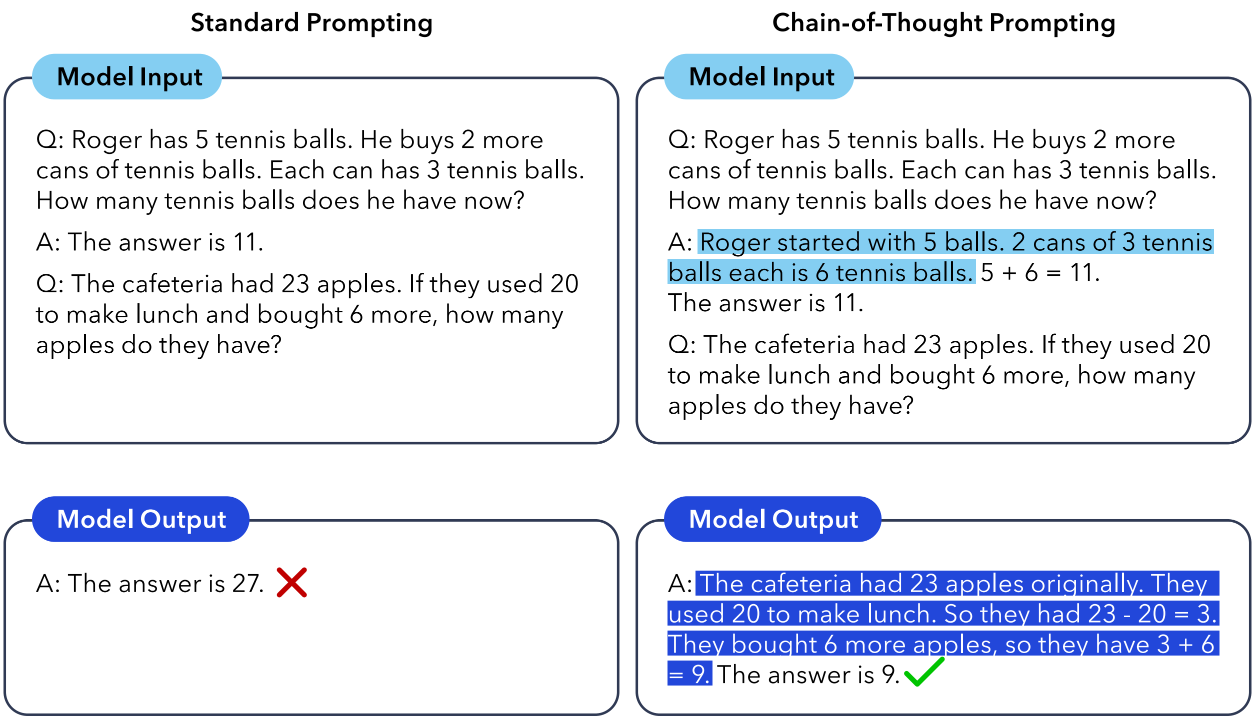


Figure 1. Standard prompting vs. CoT prompting (source: Wei et al. 2023)

Chain-of-thought is therefore crucial for LLMs because it enables the model to demonstrate a clear, coherent thought process and provides interpretable insights into the model’s behavior - thus making it easier to debug. In a recent study, the Google Research team found that chain-of-thought prompting substantially outperformed standard reasoning with a solve rate of 57% compared to 18% in GSM8K, a benchmarking dataset of elementary math word problems (source: Wei et al. 2023).

How to train an LLM to perform chain-of-thought reasoning

As with all LLM training, high-quality data is essential to consistently producing accurate results. Supervised fine tuning is an effective methodology for refining large language models by training the existing model on pre-labelled data sets. This method is ideal for training an LLM to perform chain-of-thought reasoning because the labelled data sets provide explicit input-output pairs for the model to replicate in future outputs. Supervised fine tuning also increases the efficacy of chain-of-thought reasoning as training the model on pre-labelled data sets ensures your data has previously been evaluated for quality assurance.

The subsequent case study explores the process of developing a high-quality math dataset to perform supervised fine tuning and refine the mathematical reasoning capabilities of AI.

The Challenges of Chain-of-Thought Reasoning

While studies indicate that CoT can enhance reasoning, LLMs may still generate incorrect final answers despite having correct reasoning or provide a correct final answer but with incorrect reasoning.

CHALLENGE #1: LACK OF EXTERNAL KNOWLEDGE

Problem: Language models use static internal intelligence to generate CoT which is not grounded in external knowledge. This can lead to hallucinations during reasoning.

Solution: To overcome this challenge, the ReAct prompting strategy synergizes verbal "reasoning" with interactive "action" in decision-making tasks (see Figure 2). This method works by consulting external sources (e.g., Wikipedia) and incorporating that information into the reasoning process. The ReAct prompting approach helps to ground the model's reasoning in real-world knowledge, thereby reducing the likelihood of hallucinations and improving the accuracy of the final answers.

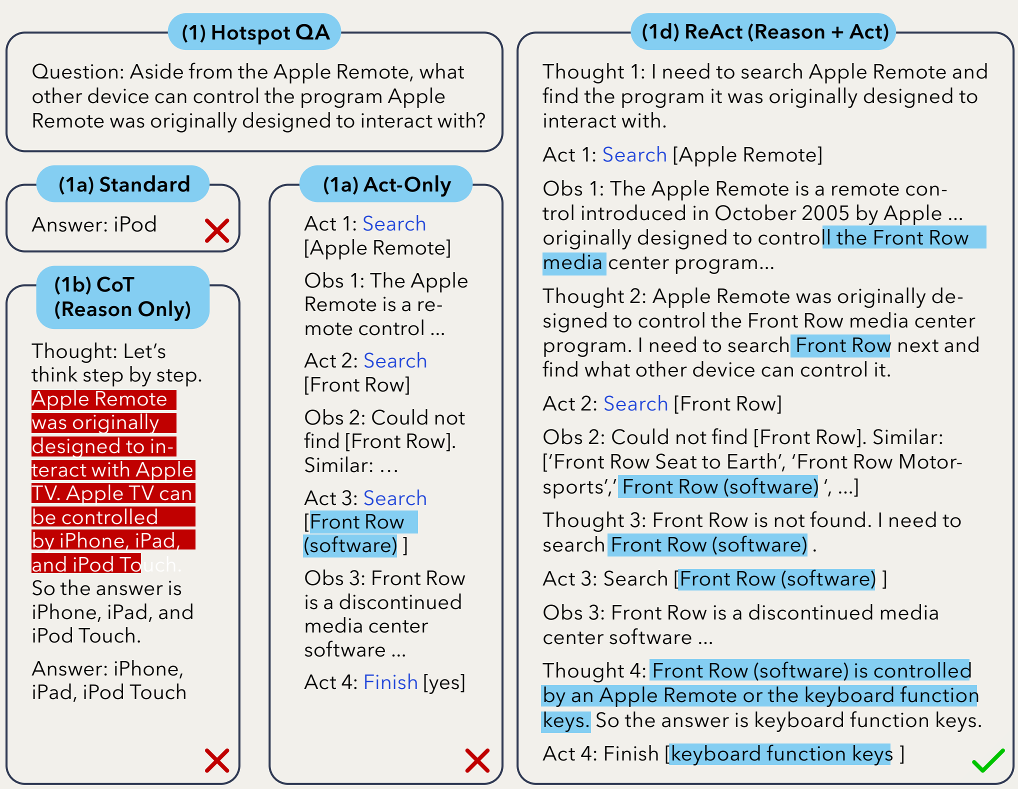


Figure 1. Standard prompting vs. CoT prompting (source: Wei et al. 2023)

CHALLENGE #2: LOWER ACCURACY ON COMPLEX TASKS

Problem: Studies have also shown that standard CoT prompting has lower accuracy on complex tasks, such as symbolic reasoning and advanced math problems, which require multiple reasoning steps.

Solution: Inspired by educational psychology, the Least-to-Most prompting technique breaks down complex question into sub-questions that are solved sequentially (see Figure 3). This method helps the model tackle complicated problems more effectively by addressing each component step-by-step, improving overall accuracy and reasoning capabilities. Experiments show that Least-to-Most prompting outperforms standard CoT prompting, especially in problems that need at least 5 steps to solve (See Figure 4).

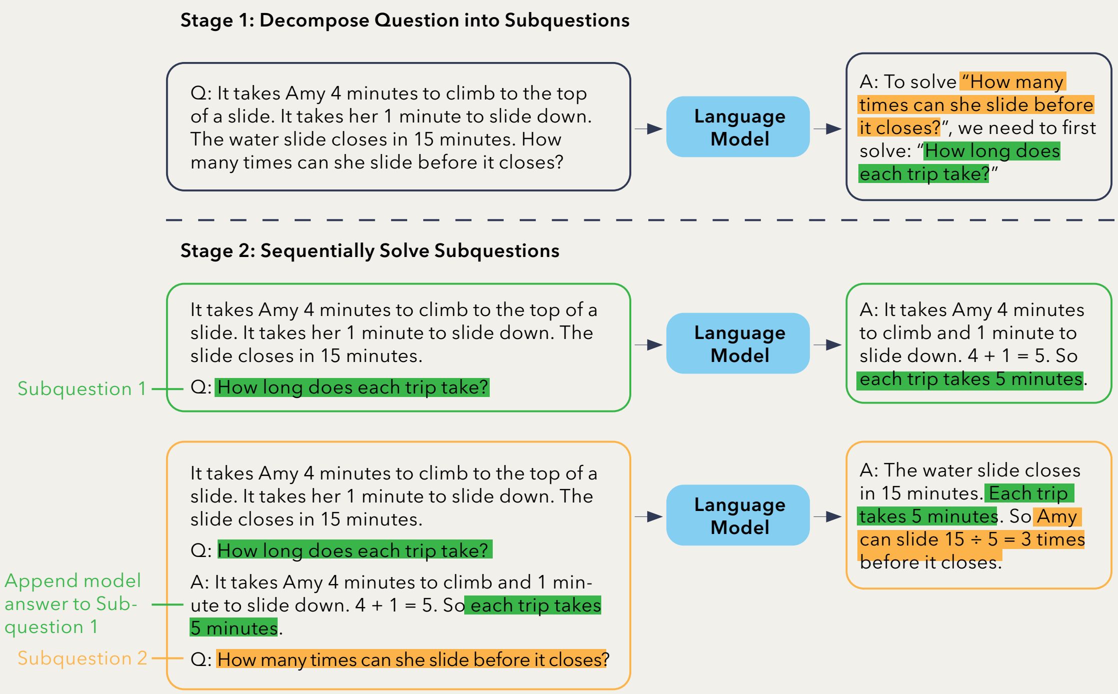


Figure 3. An example of Least-to-Most prompting (source: Zhou et al. 2023)

Accuracy by Steps (GSM8K)	All	2 Steps	3 Steps	4 Steps	≥5Steps
Least-to-Most	**62.39**	74.53	**68.91**	**59.73**	**45.23**
Chain-of-Thought	60.87	**76.68**	67.29	59.39	39.07

Figure 3. An example of Least-to-Most prompting (source: Zhou et al. 2023)

CHALLENGE #3: ERRORS IN LLM REASONING

Problem: Although CoT captures the logical progression of thoughts, the outputs can still exhibit hallucinations – such as incorrect or missing steps, flawed logic, and inaccurate calculations.

For example, we picked a question from Appen's HGMR math dataset (for further details, refer to the "Case Study" section) and sent it to Titan and Llama 2 via AWS Bedrock (see Table 1). Neither of the outputs contained the correct answer. In Titan's response, faulty reasoning began in Step 2, where it incorrectly assumed that the time spent on two distinct distances was equal. Similarly, Llama 2's response contained an error in the final few steps, incorrectly setting the total distance as x instead of 2x, leading to an incorrect final answer.

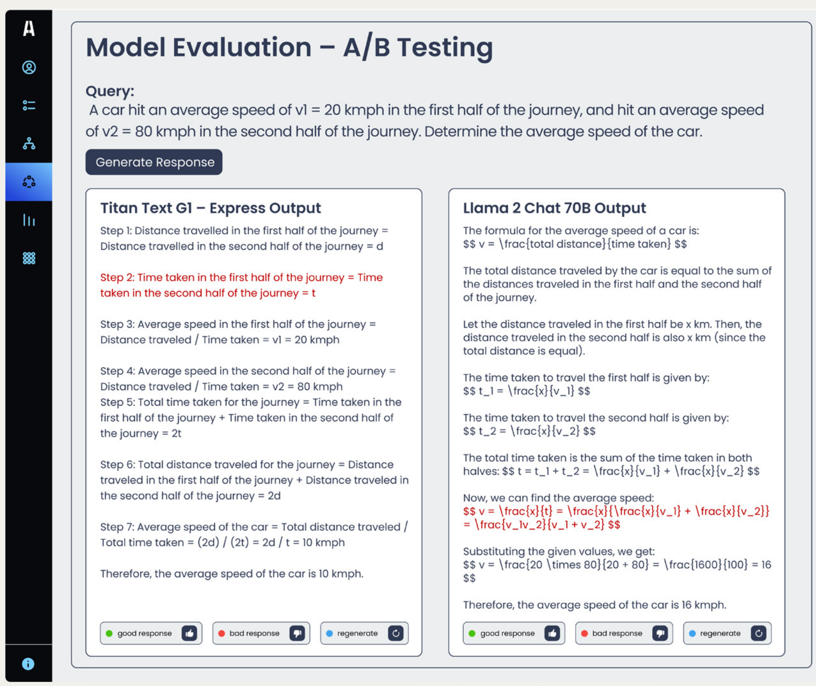


Table 1. Example of incorrect CoT reasoning

Solution: While this example demonstrates that the ability to explain and predict step-by-step thoughts needs further improvement, these errors can be easily corrected, ensuring each step is accurate and logically sound, with human-in-the-loop collaboration.

Why high-quality data is essential to CoT reasoning

High-quality CoT data is crucial for developing a language model with robust reasoning capabilities. For accurate CoT reasoning, training data should decompose the preconditions and conclusions for a given logical reasoning process and demonstrate the correctness of the logic.

How Appen delivers high-quality CoT reasoning data

As it becomes increasingly difficult to find readily usable data (copyright-free, of good quality, relevant to one's use case, etc.), turning to expert companies to craft adequate data sets is an alternative option to accelerate AI development. Appen has over two decades of experience building customized high-quality datasets for technology companies and their diverse needs and industry use cases, including chain-of-thought reasoning.

The following case study explores how Appen's expert team designed precise annotation guidelines, identified subject-matter experts, and produced valuable data for supervised fine tuning of CoT reasoning for a mathematical use case. After the initial chain-of-thought reasoning for a math problem, this data can be forwarded to experienced reviewers at Appen. These reviewers can then score, edit, and make final decisions at both the step-level and the overall response level.

Appen's annotation tool brings human experts into the process by enabling annotators to edit and score the original text and further parse the detailed attributes of each piece of text. Additionally, the Appen AI Data Platform (ADAP) excels at onboarding domain expert annotators at scale, giving task designers the freedom to create tasks that can be completed individually or collaboratively by multiple annotators.

Case Study: Human Grounded Math Dataset 1.0

This case study showcases the methods required to bring such datasets to reality, from task design to experts sourcing to quality process. Discover the steps to build a mathematical reasoning dataset from inception to release with our team of linguistics experts, applied data scientists and solutions architects.

OVERVIEW

The Human Grounded Math Reasoning Dataset 1.0 (HGMR Dataset) consists of step-by-step solutions to math problems at the high school level (Grades 9 to 12) and at the undergraduate level (including, Discrete Math and Calculus) in US English. Each instance of the dataset is composed of the following metadata: Question, Hint, Final Answer, and Step-by-Step Solution, which includes Supporting Information and Answer for each step.

Our approach to creating the HGMR Dataset integrates advanced methodologies that enhance the reliability of generated math problem sets. Recognizing the limitations seen in traditional Chain-of-Thought (CoT) processes, where logical missteps and inaccurate calculations persist, we have implemented robust verification mechanisms. By integrating AI-generated content with a rigorous vetting process conducted by domain experts, our dataset not only captures a logical progression of thought but also minimizes errors through grounded reasoning. This commitment to precision ensures that each component of our dataset—from the formulation of questions to the arrangement of possible answers—meets the highest standards of accuracy and educational value, providing users with a uniquely reliable resource for learning and application.

To bring our HGMR Dataset to life, several challenges needed to be addressed:

1. Refining a roster of reliable Math experts capable to provide accurate answer as well as clear step-by-step explanations;
2. Defining a rigorous process to ensure we would thoroughly validate all input data;
3. Designing a human task into our ADAP platform to efficiently gather the math data set from our vetted pool of experts.

1.REFINING THE ROSTER

The first step toward building the HGMR Dataset was to identify trustworthy with a strong grasp of math at the required level, and able to convey clear step-by-step reasoning in a close to perfect English.

A two-step qualification process

To qualify these quality experts in the Appen +1 million contributors crowd, we designed a two steps approach.

The first step was to identify and recruit contributors with relevant math expertise and to categorize them into two groups: group A for high school and group B for undergrad. They were invited to take a Multiple-Choice Question (MCQ) quiz, which comprises of 12 MCQs, for which they had to mark the correct answer and provide a CoT reasoning. We used a grading threshold based on logical correctness of the answer and set at 0.90, to move contributors to the second step.

The second step consisted in Appen Math SMEs reviewing the CoT demonstration provided by the contributors. This step was integral to the process as SMEs reviewed the contributors' logical reasoning as well as natural language proficiency.

Two issues might question to confidently use MCQs to vet individual skills. The first issue is that most MCQs are publicly available with their correct answer, hence one cannot be certain that correctly answering the MCQs is a proof of domain mastery. The second issue is that creating MCQs is not only time consuming but also requires domain expertise.

An LLM-based solution for Skill Qualification

To solve this problem, we used an LLM-based solution powered by Anthropic Claude 2 and maximum variance algorithm, to generate MCQs in a reduced time frame while maintaining quality.

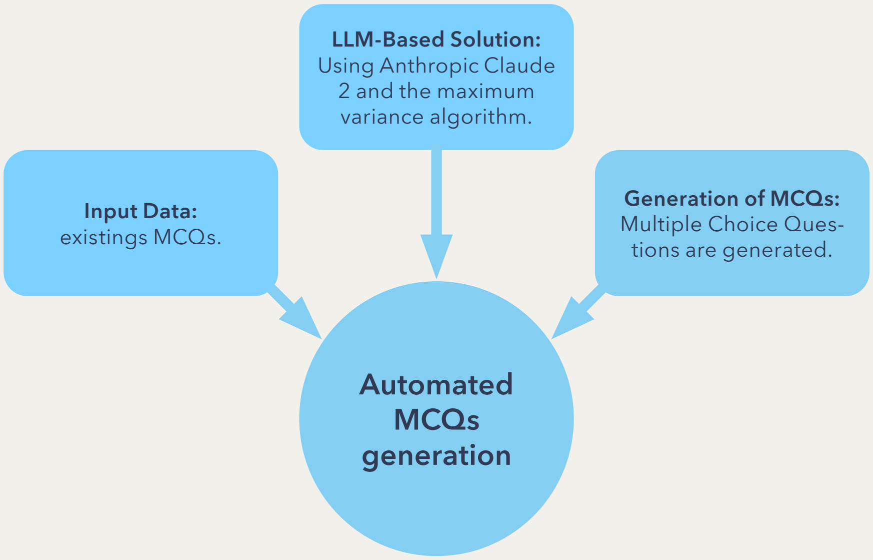


Fig.5: LLM-based solution for Skill Qualification

Only two math experts were necessary to evaluate the math prompts, responses, and choices generated automatically using the LLM-based solution. The evaluation aimed to ensure the “clarity” and “relevance” of the data. As a result, we were able to select 60 high-quality MCQs.

We picked 12 out of these 60 generated MCQs, to be distributed to over 140 math experts among the Appen crowd who completed the quiz on-line. We selected 14 top contributors, based on their proven proficiency in solving math problems and providing detailed reasoning, to join the HGMR Dataset project calibration phase.

In this final selective round, candidates who passed the quiz proceeded to project-specific calibration. The calibration phase involved setting up the real task design within the Appen SaaS platform: experts were presented with the guidelines and asked to submit a few initial responses. These responses were then graded by Appen specialists and individualized feedback was provided to each worker based on their performance.

2. DEFINING THE PROCESS

Refining the roster was crucial, but devising a strategy to enable experts to develop mathematical reasoning was equally essential, involving considerations about the types of problems to solve, the uniqueness of valid solutions, and methods for encouraging and assessing quality.

We addressed the first problem by designing a math problem generation strategy to collect and enrich problems with mathematical solutions and reasoning, and tackled the second by researching collaboration processes and bonusing strategies to incentivize high-quality math data creation.

Our math problem generation strategy

We combined AI capabilities with expert oversight to generate and refine math problem sets. These sets involve initial generation, followed by expert vetting and extensive post-processing to ensure diverse, unbiased questions and choices while preventing inaccuracies.
The output resulted in 4 elements: the question, one correct and three incorrect choices, a hint, and the answer.

Only questions that contain unambiguous phrasing, accurate hints, and correct answers were selected for inclusion in the input dataset and subsequently forwarded to the experts for solution generation.

Beyond QA, the collaborative approach

To define the most efficient process to gather faithful step-by-step reasoning for math problems at high school and undergrad level, we researched different approaches, finding inspiration in game theories, organizational behavior science and human-computer-interaction.

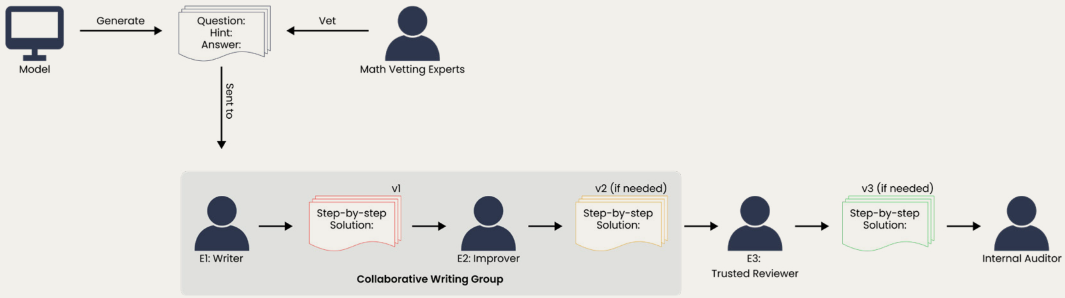


Fig. 6: HGMR Dataset creation process

In the HGMR Dataset creation process, writers (E1) and improvers (E2) collaborate on a high-quality solution, with writers generating an initial version and improvers refining it to finalization, assessed by reviewer who determines bonuses based on their performance. Consistently high performers may be promoted to reviewer roles with higher pay and additional bonuses. The reviewers' performance is regularly audited by a specialist, ensuring project quality through bonuses tied to audit results.

Our goal was to maximize cooperation between the different experts while not encouraging correction for the sake of correction. To achieve this, we created a tiered bonusing strategy that incentivizes math experts to produce high-quality data and reviews consistently. This approach ensures that our experts work together effectively to produce the best possible results.

Having the set of problems to be solved and a tailored cooperative strategy was not enough to ensure high-quality results. We needed to ensure our task design was smooth and our guidelines robust and unambiguous enough to foster efficiency.

3. DESIGNING THE TASK

We created a sequence of 3 evaluation tasks, whose combined results would lead to the most qualitative step-by-step reasoning for our selected math problems. Not only was the way these tasks were assembled important, but also the way each of these 3 tasks were specifically designed to ensure clear steps, minimal cognitive load and smooth text input was integral to the success of our dataset creation process.

Overall project architecture

Appen's AI Data Platform facilitates the creation of intricate project canvases, combining various tasks and enabling seamless data transfer between them, as configured by the task designer.

We devised a specific project to oversee the entire dataset lifecycle, from input to enrichment and correction to output. Within this project framework, data transitions automatically between the three tasks, based on completion rates and validation scores.

The project canvas starts with the initial dataset, comprising the question, a hint, and the answer. Initially, the data is presented to the writer, who inputs the step-by-step solution using our Smart Text tool enabling LaTeX formatting for math equations. Subsequently, the data flows to the improver, who enriches or rectifies as necessary. Finally, it is routed to the reviewer, who assesses whether the answer meets the expected quality standards and determines if any corrections are required. After this step, Appen Math SMEs are also able to review and audit the final output as needed.

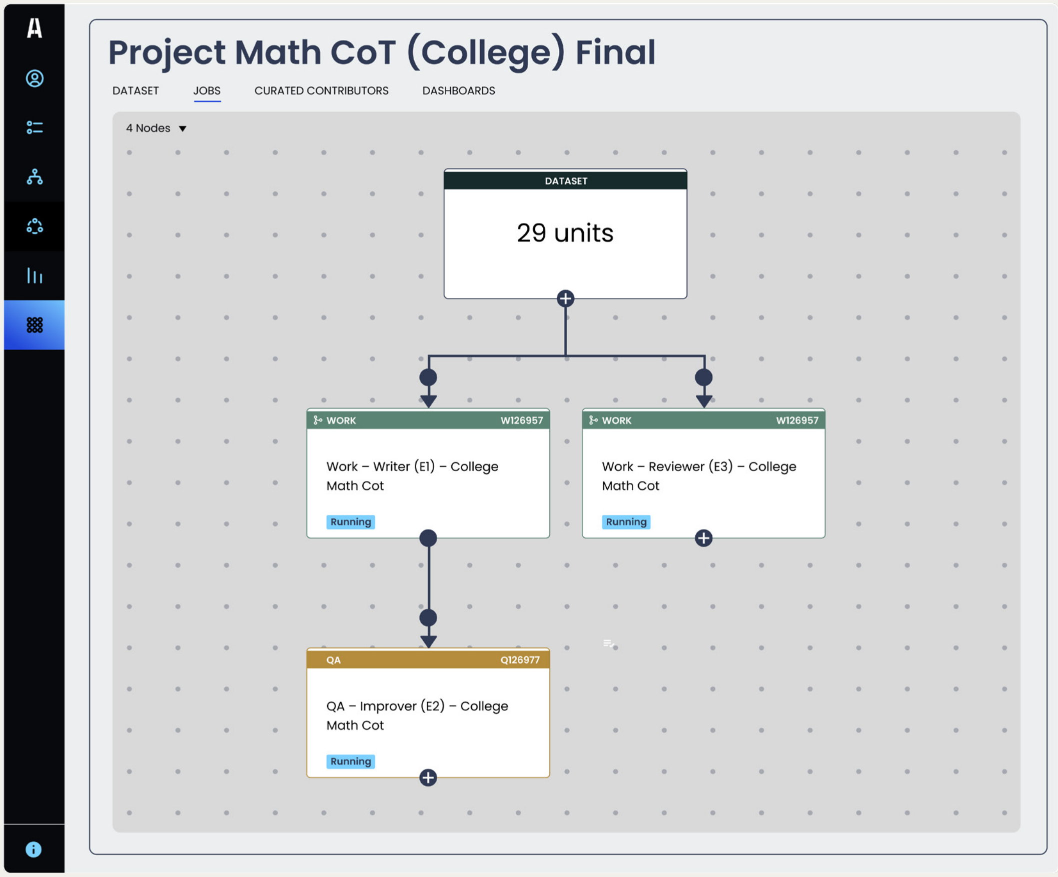


Fig. 7: HGMR project canvas in Appen's AI Data Platform

Individual task design

The layout for all tasks in our project was carefully designed to provide a seamless experience for writers, improvers and reviewers. On the left side of "task 1 preview", essential information such as the problem, final answer, and hints were displayed. This ensured contributors had clear guidance throughout. In the right side of "task 1 preview", a scrollable page allowed step-by-step reasoning, with space for contributors to input supporting evidence and logical processes. In "task 3 preview" we collected evaluator feedback on the quality of the reasoning provided. This layout prioritized clarity and ease of use, enhancing the overall user experience.

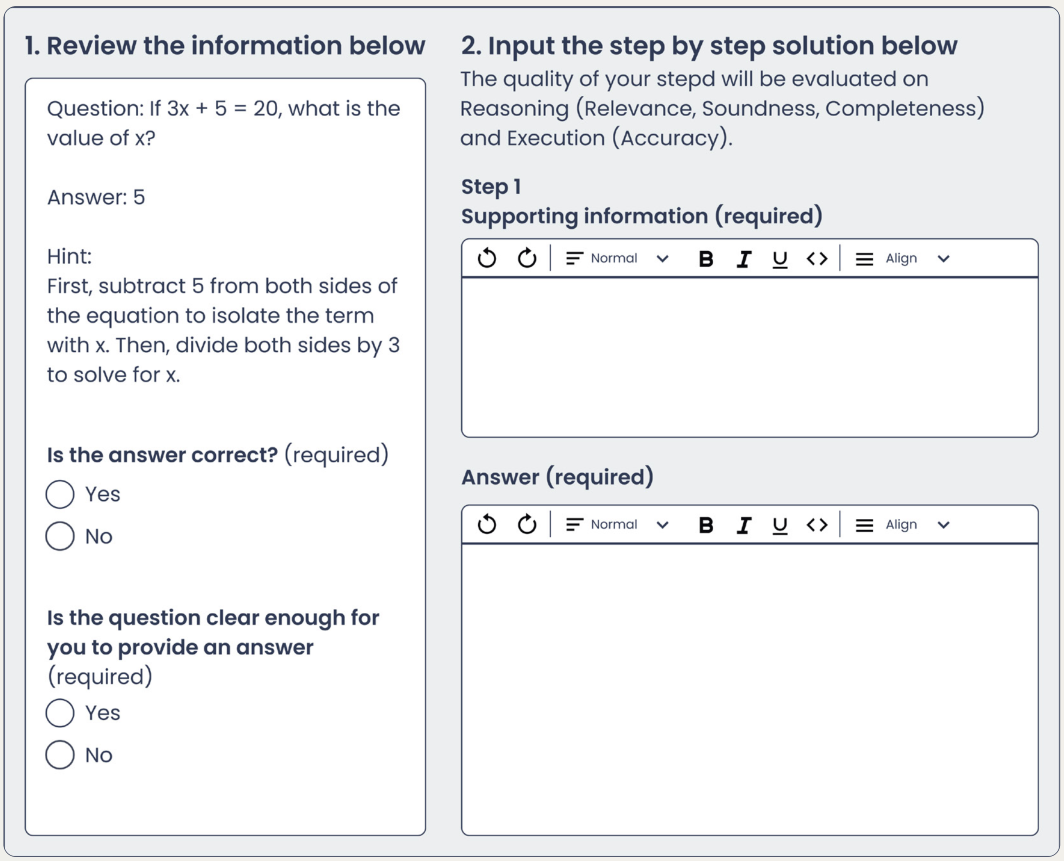


Fig 8. Task 1 preview

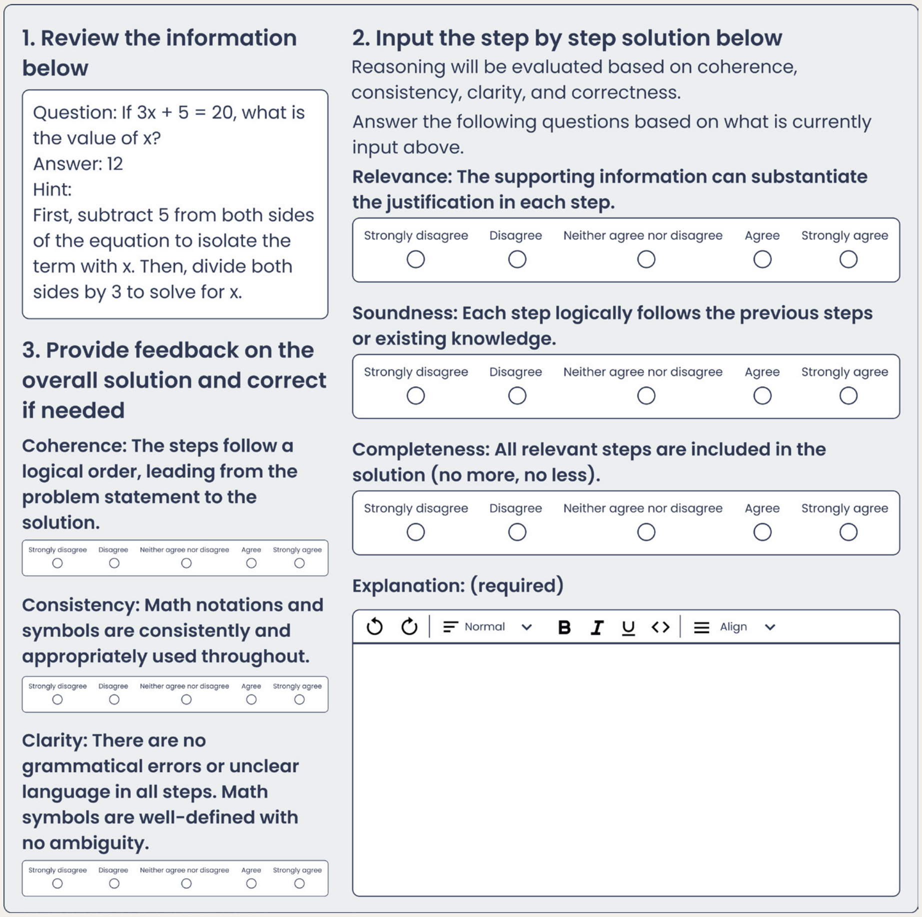


Fig 9. Task 3 preview

All tasks included ADAP native features to ease the experts' work: we implemented questions display based on conditional logic and we used our Smart Text tool, enabling diverse validators such as spelling and grammar check, but also a rich text editor allowing for using LaTeX formatting.

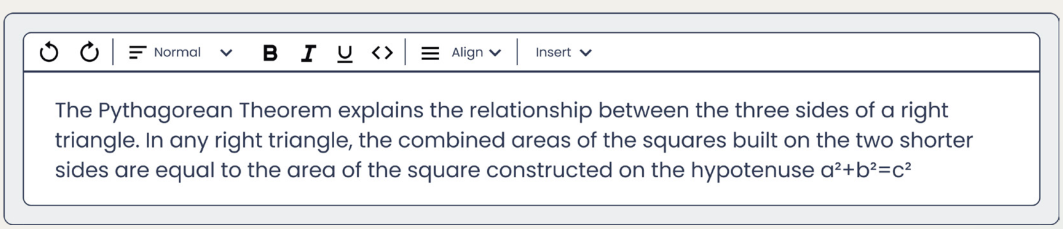


Fig 10. Smart Text tool, enabling Rich Text Editor and LaTeX formatting

Guidelines

Writing actionable guidelines is at the core of the task design. Not only the guidelines should emphasize the overall objective of the task being completed, giving the workers enough of the big picture for them to position their own contribution into it, but they also must explain in detail how to proceed, define the entire vocabulary used to describe the work to be done and cover general use cases as well as edge cases and corner cases.

Overview

The overall goal of this project is to improve the step-by-step reasoning ability of Large Language Models (LLMs) in solving math problems. By providing clear and insightful reasoning steps rooted in established knowledge, you empower LLMs to comprehend and navigate the complexities of step-by-step problem-solving, shaping the future of LLM.

In this job, you will be asked to explain step-by-step reasoning for given math problem and answer pairs.

To enhance LLM's reasoning capabilities and foster faithful responses to math questions, it is crucial to collaborate with math domain experts. Specifically, given a math question, you will need to

1. break the question into a few subquestions [step-by-step],
2. ground the reasoning with existing information or knowledge, e.g., a statement in the problem, known math principles, widely-held world knowledge, etc. [grounded reasoning].

Fig 11. Extract of instructions given to writers in task 1

For this specific project, we carefully defined concepts such as “chain-of-thought”, “step-by-step reasoning” or “supporting information”.

We clearly and transparently listed down the criteria that would be used to assess quality at the different steps of the project:

Criterion	Aspect	Description
Accuracy	Mathematical Correctness, Solution Completeness, Error Checking	Ensure mathematical content and solutions are correct, complete, and error-free.
Soundness	Math reasoning in each step	Ensure there are no logical gaps between steps and the reasoning in each step is substantiated by the supporting information, moving logically from the problem statement to the final answer.
Clarity	Problem Statement Clarity, Solution Clarity	Assess the understandability and simplicity of problem statements and solutions.
Relevance	Educational Alignment	Evaluate alignment with educational objectives and real-world applicability.
Consistency	Notational Uniformity, Terminological Consistency, Formatting and Style	Check for consistency in notation, terminology, and presentation across the dataset.
Language Quality	Grammar and Syntax, Natural Language Use, Ambiguity and Precision	Ensure texts are grammatically correct, natural, and precise without ambiguity.

Table 2: Criteria for evaluating step-level and overall quality

Conclusion

Appen utilized its experience in vetting domain experts and devising streamlined human computation tasks to deliver the HGMR Dataset. Plans are in motion to unveil an enhanced iteration of HGMR, incorporating a diverse array of math problems sourced from textbooks, coursework, worksheets, online resources, and educational software. By combining the power of both AI and human proficiency, Appen endeavors to furnish a resilient and groundbreaking approach to curating math datasets.

Contributing Authors

Madison Van Doren,
Content Marketing Manager

Madison Van Doren is the Content Marketing Manager at Appen where she specializes in AI thought leadership, SEO, and brand strategy. With over a decade of experience in linguistics research and the tech industry, Madison is uniquely positioned to translate complex AI concepts into actionable business insights. Passionate about responsible AI innovation, Madison advocates for sustainable and equitable growth in the industry through her writing and mentorship of young women in tech.

Alice Desthuilliers,
Principal Product Manager

Alice Desthuilliers is a Principal Product Manager at Appen, specializing in human-centric AI and NLP product innovation. With over two decades of experience in the tech industry, she emphasizes ethical AI practices, fostering collaboration, and enhancing human-AI interaction. Alice's holistic approach drives her ability to create comprehensive, user-friendly AI solutions while advocating for the importance of human oversight in AI development.

Si Chen,
Head of Strategy & Marketing

Si is Head of Strategy at Appen and has extensive experience in AI applications and AI data solutions, with a focus on generative AI. Prior to joining Appen, Si was Head of Strategy, Partnerships, and Operations at Tencent AI Lab, where she led multimodal AI projects for digital humans, gaming and content applications. She is passionate about applying cutting-edge technology and led the development of AI solutions for agriculture, energy, and healthcare industries as part of Tencent's AI for Good initiatives.

Shambhavi Srivastava,
Senior Machine Learning Engineer

Shambhavi is a seasoned Machine Learning Engineer and AI Solution Architect at Appen, specializing in creating and deploying innovative machine learning solutions such as advanced NLP models, automated machine learning pipelines, and cutting-edge computer vision applications. With over a decade of experience, she is known for leading large-scale projects that harness the power of artificial intelligence to drive impactful business outcomes.

Lu Lu,
Senior Linguist

Lu Lu is a senior linguist at Appen, specializing in creating methodologies and frameworks to improve dataset quality. She analyzes datasets and model outputs quantitatively and qualitatively, assisting clients in achieving their data objectives. She is also passionate about innovating methodologies to optimize the annotation process.

About Appen

Appen provides accurate and reliable human annotated datasets that fuel AI and machine learning for some of the world's biggest brands. With more than 25 years of industry knowledge, Appen powers many of the AI interactions we experience every day.

Our crowd solutions and expertise empower businesses to achieve their AI goals and make a significant impact in their industry.

Join us as we shape the future of AI and unlock its limitless possibilities.

References

Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I., & Sachan, M. (2023). MATHDIAL: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. https://arxiv.org/pdf/2305.14536

Martian, S., Martian, E., & Callison-Burch, C. (2023). Improving Mathematics Tutoring With A Code Scratchpad (pp. 20–28). Association for Computational Linguistics. https://aclanthology.org/2023.bea-1.2.pdf

Nakamoto, R., Flanagan, B., Yamauchi, T., Dai, Y., Takami, K., & Ogata, H. (2023). Enhancing Automated Scoring of Math Self-Explanation Quality Using LLM-Generated Datasets: A Semi-Supervised Approach. Computers, 12(11), 217. https://doi.org/10.3390/computers12110217

Ott, S., Hebenstreit, K., Liévin, V., Christoffer Hother, Moradi, M., Maximilian Mayrhauser, Praas, R., Winther, O., & Matthias Samwald. (2023). ThoughtSource: A central hub for large language model reasoning data. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02433-3

Rein, D., Li, B., Stickland, A., Petty, J., Pang, R., Dirani, J., Michael, J., & Bowman, S. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. https://arxiv.org/pdf/2311.12022

Saxena, Y., Chopra, S., & Tripathi, A. M. (2024, April 25). Evaluating Consistency and Reasoning Capabilities of Large Language Models. https://arxiv.org/pdf/2404.16478

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi Quoc, E., Le, V., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models Chain-of-Thought Prompting. https://arxiv.org/pdf/2201.11903

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS. https://arxiv.org/pdf/2210.03629