Can LLMs simulate the same correct solutions to free-response math problems as real students?

Yuya Asano , Diane Litman , and Erin Walker

Learning Research & Development Center, University of Pittsburgh
{yua17, dlitman, eawalker}@pitt.edu

Abstract

Large language models (LLMs) have emerged as powerful tools for developing educational systems. While previous studies have explored modeling student mistakes, a critical gap remains in understanding whether LLMs can generate correct solutions that represent student responses to free-response problems. We compare the distribution of solutions from four LLMs (one proprietary, two open-sourced general, and one open-sourced math models) with various sampling and prompting techniques and those from students teaching math problems to a conversational robot. Our study reveals discrepancies between the correct solutions produced by LLMs and by students. We discuss the practical implications of these findings for the design and evaluation of LLM-supported educational systems.

1 Introduction and related work

Large language models (LLMs) have been used to create (e.g., Chevalier et al., 2024) and evaluate (e.g., Jin et al., 2024) educational systems and simulate students (e.g., Benedetto et al., 2024). LLMs can ease authoring systems (Macina et al., 2023), increase the efficiency of testing systems (Jin et al., 2024), and provide low-cost practices without harming real students (Ma et al., 2024; Markel et al., 2023; Yue et al., 2024). However, such benefits rely on LLMs accurately reflecting real students' work. If they do not, systems may misjudge correct answers (e.g., when comparing student and LLM-generated solutions) or fail to generalize to the real world when evaluated or trained solely with LLMs (Zhang et al., 2025).

Thus, our goal is to answer the question: *Can LLMs generate the same correct solutions to free-response problems as real students?* We operationalize this goal through three novel lenses:

RQ1 How much can LLMs generate math solutions with diverse ideas, rather than variations in surface text?

RQ2 How much can LLMs cover students' correct solutions to free-response problems?

RQ3 Do LLMs' solutions have the same distribution of correct solutions as students?

We study these questions in a setting where students teach math problems to a conversational robot. Unlike prior work that measured divergence from human solutions (Lu et al., 2025; Ye et al., 2025) or accuracy without considering idea diversity (Liang et al., 2024; Yao et al., 2023; Zhou et al., 2024), we examine how diverse the core ideas in LLM solutions are and how well they represent students' diverse correct ideas. Our study focuses on:

- **Correct answers** rather than mistakes or misconceptions (Liu et al., 2025; Macina et al., 2023; Otero et al., 2025; Yue et al., 2024)
- **Free-response questions** rather than multiple-choice questions (Liu et al., 2025)
- **Real student data** rather than qualitative evaluation by teachers (Macina et al., 2023; Otero et al., 2025; Yue et al., 2024)

Thereby, we make three contributions: (1) our metric suite offers a novel lens on LLMs as student proxies, (2) we show the discrepancies between LLM and student problem solving, which is critical to the development of LLM-powered educational systems, and (3) we discuss implications for educational system design and evaluation.

2 Datasets

We obtained student solution distributions from the datasets in Asano et al. (2024), where undergraduate students taught up to nine (four with two steps) ratio problems to a conversational robot, either alone or in pairs, via speech. Paired students were encouraged to discuss their strategies before teaching. The datasets include two studies: one with 14 pairs and 12 solo students (these 12 students taught the robot alone because their partner did not show

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16348–16377
November 4-9, 2025 ©2025 Association for Computational Linguistics

Problem 1.2: Emma's friends are planning on swimming in the lake but Emma isn't waterproof. Emma decides to buy some waterproof paint so she can also go swimming. Her legs have 6 square inches of surface area, which requires 2 fluid ounces of paint. She also knows she needs 1 fluid ounce of paint to cover 3 square inches of her arms. If her body needs 3 fluid ounces of paint, how big is the surface area of her body?

Annotation	Description/Example	Students
1:3 (times 3)	Now that you have the ratio, you can multiply it by three to get your answer. (use the previous step to justify multiplying by 3).	.784
6:2	If you divide six by two, that's three. Since three is in the second column, multiply it by the ratio to get nine. (use 6:2 to infer that Emma should multiply by three).	.162
Surface area / paint = 3	Let's pretend that the surface area of the body is X. X divided by three would equal three, which is the volume of paint, based on the ratio.	.054
1/3 fl oz per square inch	Since we know the amount of paint needed per square inch for her arms is 1/3 fluid ounce per square inch, we can find the surface area of her body by dividing the total paint needed by the paint required per square inch.	0.0
Increment	We need 3 fluid ounces. We have 2 fluid ounces that cover 6 square inches. We need 1 more fluid ounce. 1 fluid ounce covers 3 square inches. Therefore, the total coverage is 6 + 3 = 9 square inches.	0.0

Table 1: Problem 1 Step 2 description, solution annotation options, and student distribution.

Speaker	Utterance
Robot	Can you give me a hint to figure out the surface area of body?
Student A	Okay. So it would be nine, right?
Student B	Yeah. Times three.
Student A	So should we say use the same ratio in step one to figure out the ratio in step two?
Student B	Okay.
Student A	Emma, use the same ratio you used in step one to figure out step two.
Robot	I get it. What's next?
Student A	Okay girl.
Student B	This area is like three times standard volume of paint.
Student A	Maybe we should just like outright tell her multiply the volume of paint by the ratio?
Student B	Yeah.
Student A	Emma, multiply the volume of paint needed for your body by the ratio we found in step one.
Robot	Okay. So we know we have three ounces more paint. What do we do with that though?

Table 2: An example dialogue between students and a robot (Emma) in Problem 1 Step 2 (Table 1). This dialogue is annotated as 1:3 (times 3) because of "use the same ratio" and "multiply ... by the ratio."

up in the study), and another with 15 pairs (2 excluded from this paper because they paired up with researchers due to no-show partners), totaling 39 dialogues and 404 solutions.

We annotated students' first solutions to each problem/step based on the underlying formulas. One author developed the annotation scheme during his annotation and trained annotators to get two annotations per problem/step (avg. Cohen's $\kappa = 0.491$, moderate agreement (Landis and Koch, 1977)). The value of κ reflected the ambiguity in student solutions (see Tables 1 and 2 for an example; Appendix A for all other problems and annotations). Thus, the author who made the annotation scheme resolved all disagreements instead of defaulting to his own annotation. Figure 1 shows the number of solutions per problem.

3 Experiments

3.1 Generation of solutions with LLMs

To generate diverse solutions to our math problems, we tested existing sampling and prompting methods designed to diversify LLMs' reasoning and make them creative. All prompts and hyperparameters are in Appendix B. We used `gpt-4o-mini`[1] as a general, closed-sourced LLM and Qwen2.5-7B-Instruct (Yang et al., 2024a) and Gemma 3 12B (Kamath et al., 2025) as general, open-sourced LLMs. Although we also tested Qwen2.5-Math-7B-Instruct (Yang et al., 2024b) as a specialized, open-sourced LLM, it did not follow instructions potentially due to catas-

[1]https://platform.openai.com/docs/models

trophic forgetting (French, 1999).[2] Altogether, we generated 5838 solutions.

Temperature sampling Macina et al. (2023) controlled temperature (t) of LLMs to simulate diverse student mistakes in math problems. We got 10 solutions for each $t \in \{0.3, 1\}$ with a Chain-of-Thought (CoT) prompt (Kojima et al., 2022).

Multi-turn reasoning Pal Chowdhury et al. (2024) obtained a tree of solution steps from an LLM to structure the interaction with their tutoring system. We tested the methods that structure LLMs' answers in multi-turn reasoning for creativity and diversity: generating multiple next steps for each step (Tree of Thoughts, **ToT**) (Yao et al., 2023), iteratively adding constraints (**denial** prompting) (Lu et al., 2025), and **paraphrasing** problems. We performed three variants of paraphrasing with LLMs: **simple** paraphrasing (Zhou et al., 2024), **translating** into other languages and cultures (see Appendix B.2.3 for which languages we used), and **replacing** numbers. We instructed LLMs to solve the modified problems with a CoT prompt and the original problem with the same method as the modified problems. We did 10 iterations for denial prompting and paraphrasing and had a maximum depth of six for ToT.

Multi-agent Liang et al. (2024) have found that LLMs cannot generate distinct solutions once they establish confidence and thus proposed a Multi-Agent Debate (MAD) framework. Inspired by their work, we instructed two instances of LLMs to introduce a new solution and critique each other. This setting is similar to classroom discussions, which have been an area of research to simulate with LLMs (Liu et al., 2024; Yue et al., 2024; Zhang et al., 2024). We provided them with the definition of different solutions used by Ye et al. (2025) because our goal is to find as many distinct solutions as possible, rather than agree/disagree.

Multiple candidates in one turn We asked LLMs to describe as many **solutions** to a specific problem or **approaches** to ratio word problems in general (inspired by Zheng et al. (2024)) as possible. Again, we defined what makes two solutions different in our prompt (Ye et al., 2025). We did the same for the method to **decompose** a problem (Sonkar et al., 2023).

In-context learning (ICL) Proper ICL examples improve LLMs' accuracy, so we hypothesize that such examples change the distributions of LLM outputs. We included two problems in the same CoT prompt as temperature sampling. They were **random** problems from GSM8K (Cobbe et al., 2021) or **related** ones crafted by an author. We tested the cases where the related problems have the **same** approach and **different** approaches. These approaches mimicked the most popular student solutions, aiming at shifting the distribution of LLM solutions toward those of students.

[2]We included the results of the temperature sampling and in-context learning in Appendix C.

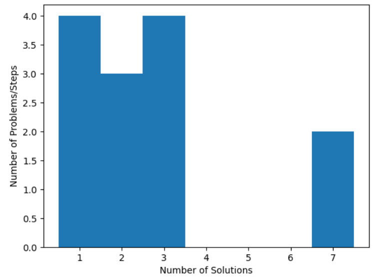

Figure 1: The number of distinct student solutions per problem/step. The overall average is 2.69, and the average of problems with multiple student solutions is 3.44.

3.2 Evaluation

One author who annotated the student dataset in Section 2 annotated all LLM solutions. We chose single annotation for LLM solutions because they tend to be clearer and less ambiguous than student ones (see examples in Appendix A vs. Appendix H). Since all RQs are about correct solutions, we discarded incorrect solutions from both students and LLMs.[3] The correctness was determined by the final answers and the intermediate steps.

RQ1 is evaluated with the number of distinct solutions generated by each method (**# slt.**). This metric is inspired by the Distinct-n score (Liu et al., 2022), but our analysis compares the mathematical ideas expressed in solutions, unlike Distinct-n, which sees only surface-level differences in wording or tokens.

RQ2 is evaluated with LLM's coverage of all (**Cov. all**) and the majority (**Cov. maj**, i.e., top k popular solutions that account for more than 50% of student solutions) student solutions.

RQ3 is evaluated with the closeness between student and LLM solution distributions (Total Variation Distance (**TVD**, Levin and Peres, 2017) and

[3]We did not investigate incorrect solutions because the accuracy of LLMs was 85% or higher for most methods (cf. Appendix D) and they have been previously explored (Liu et al., 2025; Macina et al., 2023; Otero et al., 2025; Yue et al., 2024).

the proportion of LLM solutions outside of student solution distributions (**OOD**, Lang et al., 2024). TVD is defined by

$$TVD(P_{LLM}, P_s) = \frac{1}{2}\sum_x |P_{LLM}(x) - P_s(x)|$$

where x is a solution, $P_{LLM}(x)$ is the proportion of x in the LLM's solutions, and $P_s(x)$ is the proportion of x in the students' solutions. The ranges of TVD and OOD are $[0, 1]$. See Appendix E for an example calculation.

4 Results

RQ1: Diversity of the ideas in LLM solutions The first columns of each LLM in Table 3 show that the multi-agent setting yields the most distinct correct solutions for all LLMs. In this setting, GPT4o-mini and Gemma 3 are comparable to or even better than real students (2.69 for all and 3.44 for multiple solutions, Figure 1). Asking for multiple solutions in one turn and translating problems are among the top 3 for two of the models. To see if the sampling and prompting techniques have the same effect on generating diverse solutions across different models, we looked at the correlations between the models in Table 4. There is a significant correlation between GPT4o-mini and Gemma 3 in the first row, but Qwen2.5-7B does not have any significant correlations with other LLMs.

RQ2: Coverage of student solutions No method fully covers student solutions; at least 27% of all and 8% of the majority answers are missing (Table 3). Top-performing methods vary by model: for GPT4o-mini and Gemma 3, multi-agent and multiple-candidate settings dominate; for Qwen2.5-7B, translation and related problems with different approaches perform best. The second and third columns of Table 4 verify this discrepancy between Qwen2.5-7B and other LLMs. We also looked at how the number of solutions (RQ1) is related to coverage in Table 5. All correlations are above .71 and statistically significant.

RQ3: Closeness of LLM and student distributions of solutions In Table 3, the proportion of LLM solutions outside of student solution distributions (OOD) is as low as .03, indicating that the right model and technique can keep LLMs from generating solutions different from those of students. However, the smallest Total Variation Distance (TVD) is .20, meaning, for all models and methods, there exists a solution LLMs generate at least 20% more or less likely than students.

When compared among the methods, the decomposition into subproblems ranks in the top 3 for TVD for all models. No method is in the top 3 for OOD for all models; ToT, decomposition, and ICL (except for related problems with different approaches) are in the top 3 for two LLMs. However, Table 4 shows significant correlations in OOD between the models, but not in TVD.

The correlations with the number of solutions (RQ1) are mixed. OOD has significant correlations for GPT4o-mini and Gemma 3, but not for Qwen2.5-7B. On the other hand, the correlations with TVD are significant only for Qwen2.5-7B.

5 Qualitative analysis

We qualitatively examine the strengths and limitations of top-performing methods.

Multi-agent and multiple solutions These methods produced diverse solutions with high coverage of student solutions for GPT4o-mini and Gemma 3. However, for Qwen2.5-7B, despite high solution counts, the coverage of multi-agent was not high due to repetitive summarization of a conversation history after a few turns (cf. Appendix G) in 11/13 tasks. Gemma 3 generated such summaries in the last 2.1 turns on average. GPT4o-mini did not exhibit this behavior. Occasionally, the methods generated correct answers via flawed reasoning because LLMs can see previously generated answers (cf. Appendix H).

Translate The translate method ranked highly in both diversity and coverage, except for diversity with Gemma 3 and coverage with GPT4o-mini. GPT4o-mini defaulted to unit rate reasoning in 4 out of 9 problems/steps with multiple student solutions, leading to lower coverage. Although Qwen2.5-7B and Gemma 3 had a similar trend (3 out of 9), this method outperformed others due to the lack of coverage elsewhere.

ICL with related problems We expected ICL with related problems would make the distributions of LLM solutions closer to students' by exposing LLMs to solutions similar to those of students. However, it did not always reduce TVD or OOD from temperature sampling ($t = 1$) and ICL with random problems, which shared the same CoT prompt; TVD and OOD increased when we gave different approaches to GPT4o-mini, and TVD did not change for Gemma 3. GPT4o-mini was less likely to generate the most popular student solutions (6/13) than temperature sampling (9/13).

Avg. for all problems			GPT4o-mini # slt.↑	GPT4o-mini Cov. all↑	GPT4o-mini Cov. maj↑	GPT4o-mini TVD↓	GPT4o-mini OOD↓	Qwen2.5-7B-Instruct # slt.↑	Qwen2.5-7B-Instruct Cov. all↑	Qwen2.5-7B-Instruct Cov. maj↑	Qwen2.5-7B-Instruct TVD↓	Qwen2.5-7B-Instruct OOD↓	Gemma 3 12B # slt.↑	Gemma 3 12B Cov. all↑	Gemma 3 12B Cov. maj↑	Gemma 3 12B TVD↓	Gemma 3 12B OOD↓
Temperature sampling	$t = 0.3$		1.4	.52	.58	.54	.26	1.6	.60	.81	.44	**.14**	1.0	.34	.46	.47	.22
	$t = 1$		2.0	.60	.77	.50	.22	1.8	.55	.77	.53	.25	1.2	.40	.58	.46	.23
Multi-turn reasoning	ToT		1.2	.32	.38	.63	.23	0.8	.33	.50	.45	.19	0.9	.32	.42	**.20**	**.03**
	Denial		2.1	.61	.73	.63	.35	1.8	.61	.81	.47	.21	1.2	.45	.65	.46	.21
	Paraphrase	Simple	1.8	.42	.62	.53	.28	1.3	.51	.73	.45	.23	1.4	.39	.50	.47	.15
		Replace	1.6	.39	.42	.57	.35	1.5	.45	.69	.45	.24	1.9	.49	.62	.64	.43
		Translate	2.2	.60	.77	.56	.28	2.0	**.66**	**.85**	.51	.23	1.8	.52	.69	.47	.21
Multi-agent	—	—	**3.8**	.70	.81	.60	.43	**2.1**	.53	.81	.67	.38	**4.1**	**.72**	**.88**	.58	.39
Multiple candidates	Solution		3.5	**.73**	**.92**	.54	.42	1.4	.34	.50	.50	.40	**4.1**	.66	.81	.53	.40
	Approach		1.8	.60	.81	.56	.33	1.3	.40	.62	.50	.26	1.8	.43	.54	.55	.25
	Decompose		1.2	.51	.58	**.40**	**.09**	0.4	.14	.19	**.36**	.19	1.4	.45	.62	.38	.22
ICL	Random		1.5	.52	.65	.50	.17	1.7	.61	.81	.45	.18	1.5	.49	.77	.47	.20
	Related	Same	1.5	.57	.62	.45	.15	1.5	.46	.65	.46	.21	1.3	.51	.77	.46	.15
		Different	1.5	.40	.46	.53	.24	1.9	.61	**.85**	.46	.22	1.1	.40	.62	.46	.15

Table 3: Number of distinct LLM solutions (# slt.), LLM coverage (Cov.) of all and majority (maj) student solutions, Total Variation Distance (TVD), and the proportion of LLM solutions outside student ones (OOD). Best in bold, top-3 with underline. Results only with the problems with multiple student solutions are in Appendix F.

	GPT4o-mini & Qwen2.5-7B-Instruct	GPT4o-mini & Gemma 3 12B	Qwen2.5-7B-Instruct & Gemma 3 12B
# slt.	.46 (.101)	.92 (<.001)*	.23 (.424)
Cov. all	.10 (.721)	.72 (.003)*	-.01 (.966)
Cov. maj	.11 (.713)	.54 (.045)*	.13 (.655)
TVD	.46 (.096)	.02 (.948)	.42 (.131)
OOD	.72 (.003)*	.63 (.015)*	.69 (.007)*

Table 4: Pearson's correlations (p-values in parentheses) between two LLMs. $p < .05$ is marked with *.

	GPT4o-mini	Qwen2.5-7B-Instruct	Gemma 3 12B
Cov. all	.76 (.002)*	.89 (<.001)*	.92 (<.001)*
Cov. maj	.74 (.002)*	.92 (<.001)*	.71 (.004)*
TVD	.31 (.288)	.66 (.011)*	.53 (.052)
OOD	.78 (.001)*	.25 (.398)	.77 (.001)*

Table 5: Pearson's correlations (p-values in parentheses) between the number of distinct LLM solutions and each metric, for each LLM. $p < .05$ is marked with *.

We did not see this tendency for other models.

Decomposition and ToT Decomposition and ToT performed well on TVD and OOD metrics but produced the fewest solutions. Some models often failed to generate correct answers with these methods, likely due to difficulties following the instructions: 7 problems/steps for Qwen2.5-7B with the decomposition prompt and 6 problems/steps for Gemma 3 with ToT.

6 Discussion and conclusion

We examined whether LLMs generate correct free-response solutions similar to those of students, focusing on diversity (RQ1), coverage (RQ2), and distributional similarity (RQ3). Our findings suggest three recommendations for LLM-based educational systems:

1. **Ground student simulations in diverse solutions** (all RQs). Temperature sampling alone (Macina et al., 2023) may be insufficient to simulate diverse student ideas. Similar to the simulations of persona (Liu et al., 2024; Ma et al., 2024) and skills (Markel et al., 2023; Yue et al., 2024), prompting LLMs to adopt specific solutions can enhance realism.
2. **Decouple problem and solution generation** (RQ1 & RQ2). Generating problems and solutions together (e.g., **decompose** (Sonkar et al., 2023) and Chevalier et al. (2024); Wang et al. (2024)) often limits diversity. Structured methods like multi-agent discussions or paraphrasing via translation can improve it.
3. **Select methods by purpose** (RQ2 & RQ3). For robust evaluation, use high-coverage/OOD methods. For training novice teachers, favor low-TVD/OOD methods to simulate typical student responses.

Future work should test the generalizability of our findings across models and datasets, as techniques varied in effectiveness (Benedetto et al., 2024). Our datasets are limited in settings, domains, and student populations, and annotations may be biased since annotators could identify solution sources. In broader contexts, student solutions may be more diverse, posing greater challenges for LLMs. Our results highlight the need for new methods to better align LLM and student distributions. High coverage and low divergence may be achievable via external data on student behavior (Yue et al., 2024).

Limitations

In the dataset we use, student solutions may be biased for at least four reasons. First, all of the data comes from the University of Pittsburgh (Asano et al., 2024). Although they have some racial diversity (17 White, 13 Asian, 5 Black, 1 Latino, and 4 no answer for the first dataset and 13 Asian, 2 Black, 2 Nigerian, 10 White, and 1 Multiracial for the Second dataset before exclusion), they lack gender diversity (35 females and 5 males for the first dataset and 18 females, 9 males, and 1 non-binary person for the second dataset) and geographical diversity. In addition, the solutions in this dataset might not be representative of students in other age groups. Second, they saw ratio tables on a web application when teaching problems (Figure 2). This gave them visual aids for planning their solutions. On the other hand, we converted the information in those tables into natural language for LLMs because not all of them support visual input. This prevented LLMs from visually reasoning as students did. Third, students might have adapted to how the robot responded as the conversation progressed. In later problems, students might have learned what the robot can(not) do, so they solved problems differently from earlier problems. Finally, many of the pairs had discussions before talking to the robot and thus might have deviated from their original solutions. These limitations would only have constrained the solutions students generated, and thus, in a true open-response problem with no tables or interaction with others, students might make even more diverse solutions. Nevertheless, some of these biases represent student learning over the interaction, so future work can investigate which learning stages LLMs cannot represent well by analyzing the relationship between the problem order and the distribution differences.

The scale of our evaluation is small. We tested only four small LLMs due to the limitation on funding and GPU access (we chose not to use the full-size GPT-4o to keep the comparisons fair). The dataset has only 39 dialogues, and we sampled a small number of solutions for each method. Thus, neither student nor LLM solutions are exhaustive. Still, we were able to draw implications for the design and evaluation of LLM-based educational systems and simulations so that we can ensure the transferability between research studies and real-world practice. This paper should serve as a keystone to fill a gap between research efforts on educational applications of LLMs and the real world.

Our evaluation used only one data source. This is due to a lack of datasets that collected multiple correct solutions to free-response problems from real students. Existing datasets focus on misconceptions (Liu et al., 2025; Macina et al., 2023; Otero et al., 2025) and multiple-choice questions (Benedetto et al., 2024) or do not give the probability distribution of multiple correct solutions (Ye et al., 2025).

The reliability of the annotations should also be noted. As reported in Section 2, we reached only a moderate agreement in the student dataset (Landis and Koch, 1977). The reasons for the low agreement are analyzed in detail in Appendix A. The author who created the annotation scheme resolved all disagreements instead of defaulting to his annotations, but this could have introduced biases. In addition, the LLM solutions were single-coded. Although some solutions were annotated after the resolution of disagreements in the student datasets, many were done concurrently.

Some of the methods in this paper may not have performed well because we did not do anything beyond ICL to enforce alignment between LLMs and real students. Yue et al. (2024) provided LLMs with common mistakes from the Mathematics Assessment Project (2015) to generate student-like mistakes in math problems. Although our results with ICL imply that the alignment may not always happen by giving LLMs student-like examples, more extensive evaluation beyond ICL will be beneficial.

Ethical considerations

We would like to clarify two aspects of ethical concerns. First, the datasets in Section 2 were collected after approval from a local ethics board (Asano et al., 2024) and shared with the annotators after anonymization. Second, no student data was sent to LLMs in our experiments. Instead, we only gave LLMs math problems and annotated their solutions manually.

Acknowledgments

We would like to thank the PETAL lab members at the University of Pittsburgh and anonymous reviewers for their thoughtful comments on this paper. We would also like to thank Timothy Nokes-Malach and Adriana Kovashka for supporting the

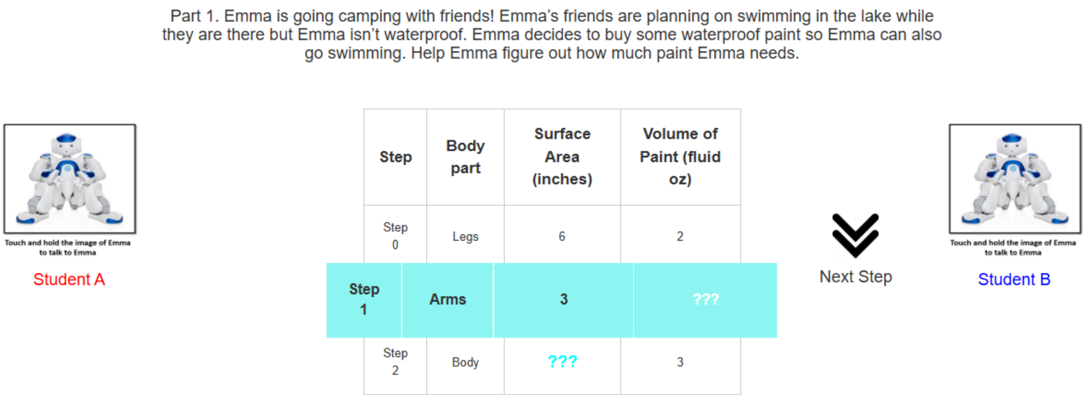

Figure 2: A screenshot of the web application.

entire project as co-PIs and undergraduate annotators (Jude Gilligan, Krit Ravichander, Travis Mindel, and Zeana El-Hajomar). This work was supported by Grant No. 2024645 from the National Science Foundation, Grant No. 220020483 from the James S. McDonnell Foundation, and a University of Pittsburgh Learning Research and Development Center award. The authors used ChatGPT to proofread author-generated text in this paper and GitHub Copilot to write the code to run LLMs for inference and analyze data (e.g., computing numbers in Table 3 and plotting the histogram in Figure 1).

References

Yuya Asano, Diane Litman, Quentin King-Shepard, Tristan Maidment, Tyree Langley, Teresa Davison, Timothy Nokes-Malach, Adriana Kovashka, and Erin Walker. 2024. What metrics of participation balance predict outcomes of collaborative learning with a robot? In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 575–583.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Aragon, Arturo Rodriguez Fanlo, Simon Frieder, Simon Machado, et al. 2024. Language models as science tutors. In *International Conference on Machine Learning*, pages 8310–8335. PMLR.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Patrik Haslum, Nir Lipovetzky, Daniele Magazzeni, and Christian Muise. 2019. An introduction to the planning domain definition language.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2024. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. *arXiv preprint arXiv:2410.04078*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hao Lang, Yinhe Zheng, Yixuan Li, Jian SUN, Fei Huang, and Yongbin Li. 2024. A survey on out-of-distribution detection in NLP. *Transactions on Machine Learning Research*.

David A Levin and Yuval Peres. 2017. *Markov chains and mixing times*, volume 107. American Mathematical Soc.

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.

Naiming Liu, Shashank Sonkar, and Richard G Baraniuk. 2025. Do llms make mistakes like students? exploring natural alignment between language models and human error patterns. *arXiv preprint arXiv:2502.15140*.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Dublin, Ireland. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2025. Benchmarking language model creativity: A case study on code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2776–2794, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiping Ma, Shiyu Hu, Xuchen Li, Yipei Wang, Shiqing Liu, and Kang Hao Cheong. 2024. Students rather than experts: A new ai for education pipeline to model more human-like and personalised early adolescences. *arXiv preprint arXiv:2410.15701*.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Mathematics Assessment Project. 2015. Mathematics assessment project. `https://www.map.mathshell.org`. Accessed in September 2025.

Nancy Otero, Stefania Druga, and Andrew Lan. 2025. A benchmark for math misconceptions: bridging gaps in middle school algebra with ai-supported instruction. *Discover Education*, 4(1):277.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. 2025. Assessing the creativity of llms in proposing novel solutions to mathematical problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25687–25696.

Murong Yue, Wenhan Lyu, Wijdane Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2024. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *arXiv preprint arXiv:2404.06711*.

Long Zhang, Meng Zhang, Wei Lin Wang, and Yu Luo. 2025. Simulation as reality? the effectiveness of llm-generated data in open-ended question assessment. *arXiv preprint arXiv:2502.06371*.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

Yue Zhou, Yada Zhu, Diego Antognini, Yoon Kim, and Yang Zhang. 2024. Paraphrase and solve: Exploring and exploiting the impact of surface form on mathematical reasoning in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2793–2804, Mexico City, Mexico. Association for Computational Linguistics.

A Problems and annotations

Tables 1 and 6-17 include the problems and annotation schemes we gave to annotators (note that Problem 1 Step 1 was used to scaffold students with researchers). The annotators were undergraduate students at the University of Pittsburgh. They consisted of 2 males and 2 females and got paid $12/hour. We explained to them the purpose of this research and told them that they would be acknowledged once a paper was accepted, either verbally or in an email. We omitted annotations for incorrect solutions for the purpose of this paper. The distributions of student solutions were normalized after discarding incorrect solutions. The student distributions in the Tables may not sum to 1 because of rounding.

There are two potential causes of the low agreement in Section 3.2. First, the student solutions were often ambiguous in conversations. For example, two annotators disagreed on the dialogue in Table 2. One annotated as 1:3 (times 3), while the other annotated as Surface area/paint = 3), possibly because their first turn addressed to the robot "use the same ratio you used in step one to figure out step two" was ambiguous about the students' idea. Second, some annotators could not find the start of the problem because some technical issues during the teaching sessions resulted in changes in the problem order or skipped problems. If one of the annotators missed a problem, it was treated as a disagreement.

B Prompts and hyperparameters

This section describes all prompts and hyperparameters for LLMs. We set the temperature to 0 unless otherwise stated. We did not do any top-p (Holtzman et al., 2020) or top-k (Fan et al., 2018) sampling for non-zero temperature. When a system prompt included a JSON format for the output of Qwen2.5-7B-Instruct and Gemma 3 12B, we used an equivalent JSON schema for GPT4o-mini instead.

Our set of techniques covers most of the prompt construction, topological variants, and enhancement of generalized CoT discussed in Chu et al. (2024), with a few exceptions:

- We did not use tool-integrated reasoning, such as Program of Thoughts (Chen et al., 2023) and the Planning Domain Definition Language (Haslum et al., 2019), because student solutions are expressed only in natural

Problem 2.1: Emma wants to bring food and have a mini party by the lake. She just found out that 3 times as many people are coming camping as she thought! She originally needed 4 cups of juice and 8 cups of seltzer. Now she needs 12 cups of juice. How much seltzer does she need now?

Annotation	Description/Example	Students
Times 3	Since three times more people are coming, you need to increase everything that you have by three times.	.542
12/4	The ratio of the new amount of juice to the original amount is: 12 cups (new) / 4 cups (original) = 3 (ratio between 4 and 12).	.375
4/12	You're going to want to write four cups of juice over 12 cups of juice. Then, use that ratio to figure out how many cups of seltzers to buy.	.083
Per person	Emma needed 8/x cups of seltzer per person for x people. For 3x people, we can calculate the new amount of seltzer needed: (8/x) * (3x) = 24 cups of seltzer.	0.0
seltzer:people	Originally, Emma needed 8 cups of seltzer for x people. We can set up a similar proportion for seltzer: $\frac{8 \text{ cups of seltzer}}{x} = \frac{y \text{ cups of seltzer}}{3x}$ where y is the new amount of seltzer needed.	0.0
juice:seltzer	Since the ratio of seltzer to juice remains the same (8 cups seltzer for 4 cups juice), we can set up a proportion: (8 cups seltzer)/(4 cups juice) = (y cups seltzer)/(12 cups juice).	0.0
Total cups	The original total amount of liquid (juice + seltzer) was 4 + 8 = 12 cups. Now, with 3 times as many people, the total amount needed is 3 * 12 = 36 cups. Since she needs 12 cups of juice, the remaining amount for seltzer is 36 - 12 = 24 cups of seltzer.	0.0
Increment	Start with the original quantities: - Original juice: 4 cups; Original seltzer: 8 cups Incrementally add the required quantities to reach the new total: - First increment: Add 4 cups of juice and 8 cups of seltzer (total: 8 cups of juice, 16 cups of seltzer). - Second increment: Add another 4 cups of juice and 8 cups of seltzer (total: 12 cups of juice, 24 cups of seltzer).	0.0
Graph	Let the x-axis represent the amount of juice (in cups) and the y-axis represent the amount of seltzer (in cups).	0.0

Table 6: Problem 2 Step 1 description, annotation, and student distribution.

Problem 2.2: Emma wants to bring food and have a mini party by the lake. She just found out that 3 times as many people are coming camping as she thought! She originally needed 4 cups of juice and 15 hotdogs. Now she needs 12 cups of juice. How many hotdogs does she need now?

Annotation	Description/Example	Students
Times 3	Since we have three times more people, we need to multiply 15 hotdogs by three.	1.0
12/4	The ratio of the new amount of juice to the original amount is: 12 cups (new) / 4 cups (original) = 3 (ratio between 4 and 12).	0.0
Per person	Emma originally needed 15 hotdogs for 'x' people. The amount of hotdogs needed per person is: $\frac{15 \text{ hotdogs}}{x \text{ people}} = \frac{15}{x}$ hotdogs per person. For '3x' people, the total hotdogs needed is: $3x \times \frac{15}{x} = 45$ hotdogs.	0.0
hotdog:people	The ratio of hotdogs needed would maintain the same proportion: Original hotdogs ratio: 15 hotdogs / x people = y hotdogs / 3x people.	0.0
juice:hotdog	If she now requires 12 cups of juice, we can set up a proportion to determine the new requirement for hotdogs: (15 hotdogs / 4 cups of juice) = (x hotdogs / 12 cups of juice).	0.0
hotdogs per cup	Each cup of juice originally corresponds to 15 hotdogs / 4 cups = 3.75 hotdogs per cup.	0.0
15 + 30	Since the number of people tripled, the amount of hotdogs also increased by a factor of 3. Therefore, the increase in hotdogs is 15 * 2 = 30.	0.0
Graph	We can visualize the relationship between the number of people and the number of hotdogs needed by plotting a graph. Let the x-axis represent the number of people and the y-axis represent the number of hotdogs.	0.0

Table 7: Problem 2 Step 2 description, annotation, and student distribution.

Problem 3: Emma knows that she can buy 15 hotdogs for $32.25, but now she needs to buy 45 hotdogs for her party. How much money does she need to buy 45 hotdogs? Answer in decimal.

Annotation	Description/Example	Students
45/15	We are going to have to find the ratio between fifteen and forty five and then apply that to the total cost.	.579
Dollars per hotdog	Divide $32.25 by 15 hotdogs to figure out how much each hotdog costs. Then multiply it by 45.	.421
hotdog:cost	The ratio of hotdogs remains proportional, so we can write the equation: $\frac{15 \text{ hotdogs}}{32.25 \text{ dollars}} = \frac{45 \text{ hotdogs}}{x}$	0.0
Increment by 15	Fill a ratio table for 30 and 45 hotdogs.	0.0
Graph	Plot the point (15, 32.25) on the graph, representing the cost of 15 hotdogs.	0.0

Table 8: Problem 3 description, annotation, and student distribution.

Problem 4.1: Emma uses 1/20 of her battery every 2/3 of an hour. How much of her battery does she use in 1 hour? Answer in fraction.

Annotation	Description/Example	Students
battery:time (divide 1/20 by 2/3)	Divide battery usage by time to get the ratio between battery usage and time (This can be explained as a unit rate, too).	.517
Cross multiply	Cross multiply the battery usage and 2/3 hour, i.e., 1/20 * 3/2, to get the ratio between battery and time.	.172
Reciprocal of 2/3	The ratio between 2/3 hour and 1 hour is the reciprocal of 2/3.	.103
Divide by 2 and multiply by 3	Divide 1/20 by 2 to get the battery usage in 1/3 of an hour, and multiply by 3 to get an hour.	.103
Convert to minutes	Convert hours to minutes to work on whole numbers.	.035
Find battery for two hours	Multiply the battery usage in 2/3 hours by three first to get how much battery is used for two hours.	.035
Find common multiple	Find the least common multiple of one over twenty (battery usage) and two over three (time), which is sixty, in order to get the ratio.	.035
Convert battery usage to percentage	Convert battery usage to percentage by multiplying it by 100 to make it a whole number.	0.0
Answer * 2/3 = 1/20	Let R be the rate of battery usage per hour. If $\frac{2}{3}$ of an hour corresponds to $\frac{1}{20}$ of the battery used, then we can write: $R \times \frac{2}{3} = \frac{1}{20}$.	0.0
Graph	Visualize the relationship between battery usage and time with a graph.	0.0

Table 9: Problem 4 Step 1 description, annotation, and student distribution.

Problem 4.2: Emma uses 1/20 of her battery every 2/3 of an hour, which means that she uses 3/40 of battery every hour. How much of her battery is used up in three hours? Answer in fraction.

Annotation	Description/Example	Students
Times 3	Multiply the unit rate of 3/40 by three hours.	.973
Cross multiply	Set up two ratios and cross multiply so 3/40 equals blank over three	.027
Additive	Add the fraction for each hour:	0.0
Convert to minute	Convert the hours to minutes.	0.0
Convert to decimal	Use a decimal approximation: 3/40 = 0.075	0.0
Convert to percent	3/40 = 7.5% of the battery is used per hour. In 3 hours, 3 * 7.5% = 22.5% is used. 22.5% = 22.5/100 = 225/1000 = 9/40.	0.0
battery:time	$\frac{1 \text{ hour}}{x} = \frac{3 \text{ hours}}{\frac{3}{40}}$	0.0
Graph	Create a bar graph where the x-axis represents time in hours (0, 1, 2, 3) and the y-axis represents battery usage.	0.0

Table 10: Problem 4 Step 2 description, annotation, and student distribution.

Problem 5: Emma uses 2/15 of battery in half an hour. How many hours does 1 battery last? Answer in fraction.

<table>
<tr><th>Annotation</th><th>Description/Example</th><th>Students</th></tr>
<tr><td>battery:time</td><td>Use the ratio between battery and time. Battery / time stays the same, so solve 2/15 / 1/2 = 1/x where x is the number of hours one battery lasts.</td><td>.279</td></tr>
<tr><td>2/15*?=1</td><td>Use the ratio between the battery. Since you have to multiply 15/2 to move from 2/15 to 1, do the same operation to 1/2.</td><td>.240</td></tr>
<tr><td>Time/battery</td><td>Since the problem asks how many hours one battery lasts, this is equivalent to the unit rate, i.e., time per battery. To compute the unit rate, you can divide time by battery.</td><td>.160</td></tr>
<tr><td>Cross multiply</td><td>The answer is 1/2*15/2 (students typically did not give good explanations for this).</td><td>.160</td></tr>
<tr><td>2/15*7+1/15=1</td><td>To get 2/15 batteries to 1 battery, you can multiply 2/15 by 7 and add 1/15. To get the number of hours, you should calculate hours 2/15 batteries last times 7 plus hours 1/15 battery lasts, which is 1/2*7 + 1/4.</td><td>.079</td></tr>
<tr><td>Reciprocal of battery per hour</td><td>The problem asks time per battery. This is the same as the reciprocal of battery per hour. Since Emma uses 2/15 batteries in 1/2 hour, you can multiply 2/15 by 2 to get the battery per hour.</td><td>.040</td></tr>
<tr><td>Decimal</td><td>Convert the fraction of battery used into a decimal and then calculate the total time in hours.</td><td>.040</td></tr>
<tr><td>1/15 lasts 1/4 hour</td><td>Since 2/15 batteries last 1/2 hour, 1/15 battery lasts 1/2 / 2=1/4 hour. To get 1 battery from 1/15 battery, you can multiply by 15. Therefore, 1 battery lasts 1/4*15=15/4 hours.</td><td>0.0</td></tr>
<tr><td>15/2 half hours</td><td>To find out how many hours 1 battery lasts, you should determine how many half hours are in 1 battery. This is the reciprocal of 2/15, which is 15/2. 15/2 half hours is equal to 15/4 hours.</td><td>0.0</td></tr>
<tr><td>Deplete remaining</td><td>After 1/2 hour, 1 - 2/15 = 13/15 of the battery remains. If 2/15 is used in 1/2 hour, then the rate is 4/15 per hour. Time to deplete 13/15 = (13/15) / (4/15) = 13/4 hours. Total time = 1/2 + 13/4 = 2/4 + 13/4 = 15/4 hours.</td><td>0.0</td></tr>
<tr><td>Per minute</td><td>To find the usage per minute, divide the fraction by 30: (2/15) / 30 = (2/15) * (1/30) = 2/450 = 1/225 of the battery per minute. Since the device uses 1/225 of the battery per minute, it will take 225 minutes to drain a full battery.</td><td>0.0</td></tr>
<tr><td>Table</td><td>Time (hours) | Battery Used (fraction)
0.5 | 2/15
1.0 | 4/15
1.5 | 6/15 = 2/5
2.0 | 8/15
2.5 | 10/15 = 2/3
3.0 | 12/15 = 4/5
3.5 | 14/15
4.0 | 16/15 > 1 (Battery depleted before 4 hours)
Since the battery is depleted at 3.5 + (1/2)*(15/15-14/15) = 3.5 + 0.5/2 = 3.5 + 0.25 = 3.75 The battery lasts 15/4 hours.</td><td>0.0</td></tr>
<tr><td>Graph</td><td>Set up a graph where the x-axis represents time in hours and the y-axis represents the remaining/used battery percentage.</td><td>0.0</td></tr>
</table>

Table 11: Problem 5 description, annotation, and student distribution.

Problem 6: Emma wants to make sure to have enough batteries for the whole trip. She knows one battery lasts for 3 and 3/4 hours. She will be gone for 2 and 3/4 days. How many batteries should she buy at the store? Round up the answer.

Annotation	Description/Example	Students
Days to hours → divide it by $3\frac{3}{4}$	Convert $2\frac{3}{4}$ days to hours and then divide that by $3\frac{3}{4}$.	1.0
Hours to days	Convert the time from hours to days and then divide $2\frac{3}{4}$ days by that number.	0.0
Multiply 3.75 by increments	Multiply 3.75 by increments (3.75*1, 3.75*2, ...) until we reach or exceed 66 hours.	0.0
Days to hours → battery:time	1 battery / 3.75 hours = x batteries / 66 hours.	0.0
Days to hours → 3.75 * (num battery) = 66 hours	Let 'x' be the number of batteries needed. So, (15/4)x = 66	0.0
Convert to minute	Convert all times to minutes.	0.0
Battery/day	24 hours/day / (3.75 hours/battery) = 6.4 batteries/day Total batteries needed: 6.4 batteries/day * 2.75 days = 17.6 batteries	0.0
48+18 hours	How many batteries are needed for 48 hours: 48 hours ÷ 3.75 hours per battery = 12.8 batteries. How many batteries are required for the remaining 18 hours: 18 hours ÷ 3.75 hours per battery = 4.8 batteries.	0,0
Graph	On a graph, plot the total hours of the trip (66 hours) on the x-axis and the number of batteries needed on the y-axis.	0.0

Table 12: Problem 6 description, annotation, and student distribution.

Problem 7.1: Emma wants to make punch for the camping trip. It takes her 3.6 minutes to make 1.2 cups of punch. She thinks she might be able to get away with about 3 and a half cups. How long will it take her to make 3.6 cups of punch? Answer in decimal.

Annotation	Description/Example	Students
punch:time	When you have 1.2 pitchers, you multiply that by three to get 3.6 minutes. So, multiply 3.6 pitchers by 3.	.476
Minutes per cup	Emma can make 1.2 pitchers of punch in 3.6 minutes. You can find the unit rate by dividing 3.6 by 1.2.	.286
Ratios between cups	We multiply 1.2 cups by three to get 3.6 pitchers. So, you can multiply 3.6 minutes by three to get the answer.	.238
Cups per minute	Calculate the rate of punch made per minute and divide 3.6 cups by that number.	0.0
Convert to seconds	Convert the time in minutes to seconds.	0.0
Cross multiply	Simplify proportion calculation: (3.6 * 3.6) / 1.2.	0.0
Graph	Plot points for (1.2, 3.6) and (3.6, y). The slope is 3.6/1.2 = 3, so y = 3 * 3.6 = 10.8 minutes.	0.0

Table 13: Problem 7 Step 1 description, annotation, and student distribution.

Problem 7.2: Emma wants to make punch for the camping trip. It takes her 3.6 minutes to make 1.2 cups of punch. She knows that this means it takes 3 minutes for her to make 1 cup of punch. How long will it take her to make 8 cups of punch? Answer in decimal.

Annotation	Description/Example	Students
Minutes per cup	Use the same unit rate from step 1. So, multiply eight pitchers by three to figure out how long it will take.	.941
3.6:8	Figure out how much more eight is by 3.6. Then, use that ratio to find the time it would take to make it.	.059
Calculate unit rate again	Calculate the unit rate using 3.6 minutes for 1.2 cups again before solving this problem.	0.0
1.2:8	If she is making 8 cups, we first find how many times 1.2 cups fit into 8 cups.	0.0
Cups:time (1.2:3.6)	Find the ratio between cups and time using 3.6 minutes for 1.2 cups again.	0.0
Cups per minute	Divide 8 cups by 1/3 cups per minute.	0.0
Cups:time (unit rate)	$\frac{1 \text{ cup}}{3 \text{ minutes}} = \frac{8 \text{ cups}}{x \text{ minutes}}$	0.0
Additive	We can recursively calculate the time for each additional cup.	0.0
Graph	Create a simple graph where the x-axis represents the number of cups of punch and the y-axis represents the time in minutes.	0.0

Table 14: Problem 7 Step 2 description, annotation, and student distribution.

Problem 8.1: Emma is going to need to head to the store to buy the waterproof paint. However, she lives in a neighborhood very far from the store. She is going to have to ride a bike! She knows she can bike 25 miles in 2.5 hours. How long does it take her to bike 30 miles? Answer in decimal.

Annotation	Description/Example	Students
Get mph	Divide the value for distance by the value for time to figure out how much distance you can go for one hour.	.639
Hrs/mile	To figure out the amount of time per distance, divide the time by the distance.	.277
Time:distance	Use the ratio between the time and the distance. We need to find a number where, if you multiply it by 10, you will get 30 miles.	.083
Miles per minute	Convert the biking speed to miles per minute.	0.0
2.5 + 0.5 = 3	Estimate the time by recognizing that 30 miles is 20% more than 25 miles. Since 2.5 hours is 100% of the time for 25 miles, 20% of 2.5 hours is 0.5 hours, so 2.5 + 0.5 = 3 hours.	0.0
Ratio between distance	Calculate the ratio between distance and multiply it by 2.5 hours: (30 / 25) * 2.5.	0.0
Graph	Plot the known point (25 miles, 2.5 hours) on a graph where the x-axis represents time (in hours) and the y-axis represents distance (in miles).	0.0

Table 15: Problem 8 Step 1 description, annotation, and student distribution.

Problem 8.2: Emma is going to need to head to the store to buy the waterproof paint. However, she lives in a neighborhood very far from the store. She is going to have to ride a bike! She knows she can bike 25 miles in 2.5 hours, which means 10 miles per hour. How far can she go if she bikes for 7.25 hours? Answer in decimal.

Annotation	Description/Example	Students
Times 10	Because we are trying to find the distance, multiply 7.25 hours by 10 mph.	1.0
Recompute mph	Refer back to 25 miles in 2.5 hours to recompute mph.	0.0
Convert to minute	Convert the biking time into minutes: 7.25 hours = 7 hours and 15 minutes = 435 minutes.	0.0
Time:distance (2.5:25)	Set up the proportion: 25 miles / 2.5 hours = x miles / 7.25 hours.	0.0
Time:distance (mph)	$\frac{10 \text{ miles}}{1 \text{ hour}} = \frac{x \text{ miles}}{7.25 \text{ hours}}$.	0.0
7 + .25 hours	She can bike for 7 hours at 10 miles/hour (70 miles) and then for the remaining 0.25 hours (15 minutes) at the same speed (2.5 miles).	0.0
1 + 6.25 hours	We can say that Emma decides to bike for 1 hour, then takes a short break and continues biking for another 6.25 hours.	0.0
2.5 + 4.75 hours	7.25 hours = 2.5 hours + 4.75 hours. Distance in 2.5 hours = 25 miles Distance in 4.75 hours = 4.75 * 10 = 47.5 miles. Total distance = 25 + 47.5 = 72.5 miles	0.0
Graph	Represent Emma's biking journey on a graph where the x-axis represents time (in hours) and the y-axis represents distance (in miles).	0.0

Table 16: Problem 8 Step 2 description, annotation, and student distribution.

Problem 9: In preparing for camping, Emma's friends Tasha and Zach have been arguing over who is better at making s'mores. Tasha has suggested that they use math to figure out who is fastest. Tasha already has a function for how quickly she can make s'mores. It is y=2x+4, meaning that it takes y minutes in total for Tasha to make x s'mores, including 4 minutes for set up. What is the slope of the function for Zach given that it takes 9 minutes in total to make 2 s'mores and it takes 1 minute for set up?

Annotation	Description/Example	Students
y=mx+b	We are using the equation Y equals MX plus B, which is helpful to understand linear problems. So it would be nine equals two times M plus B.	1.0
Making s'mores is 8 minutes	Time for making s'mores = 9 minutes - 1 minute = 8 minutes.	0.0
Graph	Using a graphical approach, we can plot the points for Zach's s'mores making function. The point (0, 1) represents the setup time, and the point (2, 9) represents the total time for making 2 s'mores.	0.0

Table 17: Problem 9 description, annotation, and student distribution.

language.
- We did not try graph structures (e.g., Besta et al. (2024)) because they are an extension of tree structures with aggregation and refinement, which do not add diversity of reasoning paths. This is also the reason why we did not test methods for verification and refinement.
- Efficiency was not our focus.

B.1 Temperature sampling

We gave the following system prompt, followed by a user prompt containing a math problem.

> **Temperature sampling**
>
> Solve the following math problem. Show your step-by-step work and state the answer at the end.

The temperature was either 0.3 or 1.0.

B.2 Multi-turn reasoning

B.2.1 Tree of Thoughts (ToT)

Yao et al. (2023) proposed two ways to generate thoughts: sampling i.i.d thoughts from CoT and proposing sequential thoughts. Since the first way is the same as the temperature sampling, we asked LLMs to generate one step at a time; to get the first step, we prompted,

> **ToT: First step**
>
> Give me the possible first steps to solve the following math problem. Please give me as many distinct ideas as possible. The first steps must contain numbers only given in the problem and have only one equation. Do not give me the second steps. Format your answer in the following JSON format:
> {"ideas": [{"explanation": "a text explanation of the first or next step", "equation": "a math equation of the step with the result of the calculation"}]}

and to get the next steps, we prompted,

> **ToT: Next steps**
>
> Give me the next step of the following incomplete solution. The next steps must contain information only given in the problem or the previous step and naturally follow the previous step. Please give me as many distinct ideas as possible. Do not go to the steps further. Do not introduce any new methods. Format your answer in the following JSON format:
> {"ideas": [{"explanation": "a text explanation of the first or next step", "equation": "a math equation of the step with the result of the calculation"}]}

We observed that the generated steps are often duplicated. Thus, we asked LLMs to remove duplicates:

> **ToT: Remove duplicates**
>
> Please remove duplicated pairs of an explanation and an equation that do the same calculation and output in the same JSON format.

Following the thought generation, ToT evaluates thoughts based on how well they are in reaching the final solution. Since our goal was not to eliminate bad thoughts, we evaluated thoughts based on whether they reach the final solution regardless of their correctness:

> **ToT: Final solution**
>
> Does the solution give an answer to the problem, regardless of its correctness? Explain and answer True or False in the following JSON format:
> {"explanation": "Your explanation", "answer": "True or False"}

B.2.2 Denial prompting

Denial prompting (Lu et al., 2025) involves two stages: response generation and technique detection. Response generation requires LLMs to generate solutions without using a specified set of techniques, which are math formulas or equations in our context. The system prompt was the same as temperature sampling (cf. Appendix B.1). We added prohibited techniques as a user prompt:

Denial prompting: Prohibiting techniques

> DO NOT use the following equations or equivalents:
> - <List of formulas or equations>
> Problem: <problem description>

During technique detection, we asked LLMs to extract formulas or equations from a solution from response generation:

Denial prompting: Technique detection

> You are a math teacher. Detect all equations from the input. Only select the ones with the equal symbol.

We asked Qwen2.5-7B-Instruct and Gemma 3 12B to output in a Python list and GPT-4o mini to output in the following JSON:

{“equations”: [{“equation”: “a math equation used in the solution”}]}

B.2.3 Paraphrase

Paraphrases were sampled with temperature = 1.0, but we used temperature = 0.0 when LLMs solved a problem.

Simple We followed Zhou et al. (2024) to paraphrase our math problem:

Paraphrase

> Paraphrase the following math problem.

After paraphrasing, we gave the same prompt as temperature sampling (cf. Appendix B.1) to solve the paraphrased problems.

Replace We extended the simple paraphrasing above to allow LLMs to create a new, similar problem by replacing numbers:

Replace: Create a new problem

> Create a new problem similar to the following problem by, for example, replacing the numbers.

After paraphrasing, we gave the same prompt as temperature sampling (cf. Appendix B.1) to solve the new problems. Finally, for each new problem, we prompted LLMs to solve the original problem similarly to the solution to the new problem. To do so, we gave the following system prompt, followed by a new problem and a solution as the user and assistant turns, respectively:

Replace: Solve the original problem

> Solve the following math problem using the same approach as the following sample problem and solution.

Translate We translated the original problems and the CoT prompt (cf. Appendix B.1) in English (US) to Spanish (Latin American), French, German, Russian, Chinese, Japanese, Arabic, Czech, Hindi, Icelandic, and Ukrainian. These languages are either official languages of the United Nations or languages used at the ninth conference on machine translation (WMT24). The translation and adaptation into other cultures were done by GPT4o-mini, Qwen2.5-7B-Instruct, and Gemma 3 12B with the following prompt:

Translate: Translation

> Translate the following math problem into <language> and modify it to adapt it to <culture> culture (for example, by changing the name of a person): <problem>

The translation of the CoT prompt was done by GPT4o to ensure the correctness. The CoT prompt in different languages is as follows:

- Spanish: Resuelve el siguiente problema de matemáticas. Muestra tu trabajo paso a paso y da la respuesta al final.
- French: Résoudre le problème mathématique suivant. Montrez votre travail étape par étape et donnez la réponse à la fin.
- German: Lösen Sie das folgende Mathematikproblem. Zeigen Sie Ihre Schritt-für-Schritt-Berechnungen und geben Sie die Antwort am Ende an.
- Russian: Решите следующую математическую задачу. Покажите вашу работу шаг за шагом и укажите ответ в конце.
- Chinese: 解决以下数学问题。展示你的逐步解题过程，并在最后给出答案。
- Japanese: 次の数学の問題を解いてください。手順を一つ一つ示し、最後に答えを記載してください。
- Arabic: حل مسألة الرياضيات التالية. اعرض عملك خطوة بخطوة وذكر الإجابة في النهاية.

- Czech: Vyřešte následující matematický úkol. Ukažte svůj postup krok za krokem a na závěr uveďte odpověď.
- Hindi: निम्नलिखित गणित समस्या को हल करें। अपने चरण दर चरण काम को दिखाएं और अंत में उत्तर बताएं।
- Icelandic: Leystu eftirfarandi stærðfræðiverkefni. Sýndu vinnu þína skref fyrir skref og gefðu út svarið í lokin.
- Ukrainian: Розв'яжіть наступну математичну задачу. Покажіть покрокове рішення та вкажіть відповідь наприкінці.

Then, each LLM solved the translated problems with the corresponding CoT prompts and translated their solutions to English. Finally, we did the same thing as replace to ask LLMs to solve the original problem in the same way as the translated problems.

B.3 Multi-agent debate

The MAD framework requires LLM agents to engage in tit-for-tat. Liang et al. (2024) did so by allowing LLMs not to fully agree with each other. However, since our setting was not about agreeing or disagreeing, we told LLMs to critique each other in a system prompt as follows:

Multi-agent

We are math teachers who are discussing a problem to come up with as many different model solutions as possible. For each turn, we critique the other's answer and propose a new solution not in our conversation.
Criteria for evaluating the difference between two mathematical solutions include:
1. If the methods used to arrive at the solutions are fundamentally different, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
2. Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the solutions can be considered different;
3. If two solutions rely on different assumptions or conditions, they are likely to be distinct;
4. A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
5. If one solution is significantly simpler or more complex than the other, they can be regarded as essentially different, even if they lead to the same result.

The criteria for evaluating the difference come from the definitions of the difference between two mathematical solutions used by Ye et al. (2025).

Next, we prompted LLMs "Do you understand?" and fed "Yes, I understand!" as their reply. Then, we gave a problem, asked one of the instances, "Give me your solution first," and gave the solution generated by one instance to another to start a conversation. Each instance had five turns, meaning that we had 10 turns in total.

B.4 Multiple candidates in one turn

B.4.1 Multiple solutions

The system prompt to generate multiple solutions at one time is the following:

Multiple solutions

Come up with as many solutions to the following math problem as possible.

We asked Qwen2.5-7B-Instruct and Gemma 3

12B to output in a Python list whose elements are solutions and GPT-4o mini to output in the following JSON:

> {“solutions”: [{“solution”: “a solution to the problem”, “answer”: “the final answer to the problem”}]}

B.4.2 Multiple approaches

This method involves two stages, similar to the replacement of numbers and translation. First, we prompted LLMs to generate different methods to solve a math word problem about ratios with example problems and solutions:

> **Multiple approaches: List approaches**
>
> List different methods to solve a math word problem about ratios. Format your response in the JSON object:
> {“approaches”: [{“name”: “the name of the approach”, “explanation”: “the explanation of the approach”, “example problem”: “an example problem”, “solution”: “a step-by-step solution to the example problem”, “answer”: “the final answer to the example problem”}]}

Second, similarly to the replacement of numbers and translation, we asked LLMs to solve our problems following the approaches from the first stage. The system prompt was as follows:

> **Multiple approaches: Apply an approach**
>
> Can you solve the following math problem using the method described below?
> <the name of an approach>: <the explanation of the approach>

After the system prompt, we gave an example problem as a user prompt, fed a solution as a reply, and showed one of our problems in a user prompt.

B.4.3 Decompose

Sonkar et al. (2023) proposed generating scaffolding and conversational datasets with LLMs to train a conversational tutoring system. A scaffolding dataset contains problems generated from a textbook, subproblems, hints, incorrect student responses, and feedback. A conversational dataset contains a mock conversation between a student and a tutor grounded on a scaffolding dataset. Since our goal is to generate solutions to our problems, we did not generate any conversational datasets and only instructed LLMs to generate subproblems to a given problem in a system prompt in the same way as generating a scaffolding dataset:

> **Decompose**
>
> Generate sets of subproblems that the following math problem can be broken into. For each subproblem, generate a hint, one incorrect student response to the subproblem, and corresponding feedback to the student. For each set, give an answer to the math problem based on the subproblems. Generate as many sets of subproblems as possible.

We added the same definition of different solutions used in the multi-agent setting (Ye et al., 2025) after the prompt above.

B.5 In-context learning (ICL)

We sampled 10 solutions for each problem with temperature=1.0.

B.5.1 Random problems from GSM8K

We randomly picked the following two problems from GSM8K (Cobbe et al., 2021) because our problems are at the grade-school level:

- Problem A
  - Problem: Phillip’s mother asked him to go to the supermarket to buy some things and gave him $95, so he spent $14 on oranges, $25 on apples, and $6 on candy. How much money does he have left?
  - Answer: Let’s think step by step.
    1. If we add everything Phillip bought, we will have: $14 + $25 + $6 = «14+25+6=45» $45.
    2. [Final solution] He spent $45, and we know that he had $95 dollars, so now we have to subtract: $95 - $45 = «95-45=50» $50.
- Problem B
  - Problem: Tim decides to do a movie marathon. The first movie is 2 hours long. The next movie is 50% longer. And the last movie is 1 hour shorter than the combined time of the previous 2 movies. How long was his movie marathon?
  - Answer: Let’s think step by step.
    1. The second movie was

2*.5=«2*.5=1» 1 hour longer than the first movie.
2. So the second movie was 2+1=«2+1=3» 3 hours long.
3. That means the first two movies had a combined time of 3+2=«3+2=5» 5 hours.
4. So the last movie had a length of 5-1=«5-1=4» 4 hours.
5. [Final answer] That means the marathon lasted 5+4=«5+4=9» 9 hours.

B.5.2 Related problems

One of the authors created problems related to our problems in Appendix A. We picked two problems for each as in-context examples as described below:

- For problems 1 and 2 (both steps), and step 2 of problems 4, 7, and 8
  - Problem: Lily is organizing a party and needs to buy balloons. She knows that two balloons last for 48 hours, which means that each balloon lasts for 24 hours. If she needs balloons for 96 hours, how many balloons should she buy to make sure they last the entire time?
  - Solution 1: Let's think step by step.
    1. First, we have to find the number of balloons needed to last 96 hours.
    2. Since each balloon lasts for 24 hours, we can divide 96 hours by 24 hours per balloon.
    3. «96 / 24 = 4»
    4. [Final solution] Thus, she needs to buy 4 balloons to last 96 hours.
  - Solution 2: Let's think step by step.
    1. We have to find the ratio between the number of balloons and the time they last.
    2. Since two balloons last for 48 hours, one balloon lasts for 24 hours.
    3. Now let's calculate how many balloons are needed for 96 hours.
    4. Since «96 / 24 = 4», we need to buy 4 balloons.
    5. [Final solution] Thus, she needs to buy 4 balloons to last 96 hours.
- For problems 1 and 2 (both steps), and step 2 of problems 4, 7, and 8
  - Problem: Jake is setting up a light show for an event. Three light bulbs last for 48 hours of continuous use. He has realized that he needs one bulb to light up for 16 hours. If he has four bulbs, how long can he light?
  - Solution 1: Let's think step by step.
    1. First, we have to find out how long one bulb lasts.
    2. Since each bulb lasts for 16 hours, four bulbs will last for «4 * 16 = 64» hours.
    3. [Final solution] Thus, Jake can light up for 64 hours with four bulbs.
  - Solution 2: Let's think step by step.
    1. We have to find the ratio between the number of bulbs and the time they last.
    2. Since three bulbs last for 48 hours, you should multiply «48 / 3 = 16» 16 by the number of bulbs to get time.
    3. Now let's calculate how long four bulbs can light up.
    4. Since «4 * 16 = 64», Jake can light up for 64 hours with four bulbs.
    5. [Final solution] Thus, Jake can light up for 64 hours with four bulbs.
- For problem 3
  - Problem: A group of friends buys 8 movie tickets for a total of $75.2. How much money will they need to buy 24 movie tickets? Answer in decimal form.
  - Solution 1: Let's think step by step.
    1. Since «24 = 8 * 3», they will buy three times more tickets.
    2. That means 24 tickets cost three times more than 8 tickets.
    3. Since the price of 8 tickets is $75.2, the total price for 24 tickets is «75.2 * 3 = »
    4. [Final solution] Thus, they will need $225.6 to buy 24 tickets.
  - Solution 2: Let's think step by step.
    1. First, we have to how much one movie ticket costs.
    2. The price of a ticket is «75.2 / 8 = 9.4» 9.3 dollars.
    3. To calculate how much money we will need to buy 24 tickets, we should multiply 24 tickets by the price of one ticket. «9.4 * 24 = 225.6».
    4. [Final solution] Thus, they will need $225.6 to buy 24 tickets.
- For problem 3
  - Problem: John can read 3.2 pages of a book in 6.4 minutes. If he needs to read

12.8 pages, how long will it take him to finish? Answer in decimal form.
  - Solution 1: Let's think step by step.
    1. We have to find the number to multiply by 3.2 pages to get 12.8 pages.
    2. «3.2 * 4 = 12.8», so we have to multiply 6.4 minutes by 4 to get how many minutes it takes to read 12.8 pages.
    3. «6.4 * 4 = 25.6».
    4. [Final solution] Thus, it takes her 25.6 minutes to read 12.8 pages.
  - Solution 2: Let's think step by step.
    1. First, we have to find the speed John reads.
    2. The speed is «6.4 / 3.2 = 2» 2 minutes per page.
    3. To calculate how much it takes for him to read 12.8 pages, we should multiply 12.8 pages by the speed. «12.8 * 2 = 25.6».
    4. [Final solution] Thus, it takes her 25.6 minutes to read 12.8 pages.
- For problem 5 and step 1 of problem 4
  - Problem: Wyatt walks 13/12 miles in 1/3 of an hour. How many hours does it take for him to walk 1 mile? Answer in fraction.
  - Solution 1: Let's think step by step.
    1. We have to find the ratio between miles and hours.
    2. The ratio is «13/12 / 1/3 = 1 / x».
    3. If we solve for x, «x=4/13».
    4. [Final solution] Thus, it takes 4/13 hours for him to walk 1 mile.
  - Solution 2: Let's think step by step.
    1. We have to find the number to multiply by 13/12 to get 1 mile.
    2. «13/12 / 12/13 = 1 / x», so we have to multiply 1/3 of an hour by 12/13 to get how many hours it takes to walk 1 mile.
    3. «1/3 * 12/13 = 4/13».
    4. [Final solution] Thus, it takes 4/13 hours for him to walk 1 mile.
- For problem 5 and step 1 of problem 4
  - Problem: Alex's laptop downloads 3/16 GB of data in a quarter of an hour. How many hours does it take to download 1 GB? Answer in fraction.
  - Solution 1: Let's think step by step.
    1. We have to find the ratio between the speed of download and hours.
    2. The ratio is «3/16 / 1/4 = 1 / x».
    3. If we solve for x, «x=4/3».
    4. [Final solution] Thus, it takes 4/3 hours to download 1 GB.
  - Solution 2: Let's think step by step.
    1. We have to find the number to multiply by 3/16 to get 1 GB.
    2. «3/16 / 16/3 = 1 / x», so we have to multiply 1/4 of an hour by 16/3 to get how many hours it takes to download 1 GB.
    3. «1/4 * 16/3 = 4/3».
    4. [Final solution] Thus, it takes 4/3 hours to download 1 GB.
- For problem 6
  - Problem: Liam wants to make sure he has enough water bottles for his hike. He knows one water bottle lasts for 4 and 1/2 hours. He will be hiking for 3 and 1/4 days. How many water bottles should he bring on the hike? Round up the answer.
  - Solution 1: Let's think step by step.
    1. First, we have to convert 3 and 1/4 days into hours.
    2. 3 and 1/4 days is equal to 3 * 24 + 6 = «3*24+6=78»78 hours.
    3. Now we can calculate how many bottles of water he will need in total.
    4. The total amount of water cleaned is «78 / (4 + 1/2) = 52/3 = 17.33»
    5. [Final solution] Thus, Liam should bring 18 water bottles for the hike.
  - Solution 2: Let's think step by step.
    1. First, we have to convert 4 and 1/2 hours to days.
    2. Since a day is 24 hours, 4 and 1/2 hours are «(4 + 1/2) / 24 = 3/16» 3/16 days.
    3. Now let's calculate the ratio between 3/16 days and 3 and 1/4 days.
    4. Since «(3 + 1/4) / (3/16) = 52/3 = 17.33», he will need 17.33 bottles.
    5. [Final solution] Thus, Liam should bring 18 water bottles for the hike.
- For problem 6
  - Problem: Sophia is planning a road trip and wants to make sure she has enough snacks for the entire drive. Each snack bag lasts for 40 and 2/3 minutes. She will be on the road for 4 and 1/2 hours. How many snack bags should she pack? Round up the answer.
  - Solution 1: Let's think step by step.

1. First, we have to convert 4 and 1/2 hours into minutes.
2. 4 and 1/2 hours is equal to 4 * 60 + 30 = «4*60+30=270»270 minutes.
3. Now, divide the total number of minutes by how long each snack bag lasts. «270 / (40 + 2/3) = 405/61 = 6.64»
5. [Final solution] Thus, Sophia should pack 7 snack bags.

  - Solution 2: Let's think step by step.
1. First, we have to convert 40 and 2/3 minutes to hours.
2. Since an hour is 60 minutes, 40 and 2/3 minutes are «(40 + 2/3) / 60 = 61/90» 61/90 hours.
3. Now let's calculate the ratio between 61/90 hours hours and 4 and 1/2 hours.
4. Since «(4 + 1/2) / (61/90) = 405/61 = 6.64», she will need 6.64 packs.
5. «2.5 * 6 = 15»
6. [Final solution] Thus, Sophia should pack 7 snack bags.

- For step 1 of problems 7 and 8
  - Problem: John can read 3.2 pages of a book in 6.4 minutes. If he needs to read 12.8 pages, how long will it take him to finish? Answer in decimal form.
  - Solution 1: Let's think step by step.
1. First, we have to find the speed John reads.
2. The speed is «6.4 / 3.2 = 2» 2 minutes per page.
3. To calculate how much it takes for him to read 12.8 pages, we should multiply 12.8 pages by the speed. «12.8 * 2 = 25.6».
4. [Final solution] Thus, it takes her 25.6 minutes to read 12.8 pages.

  - Solution 2: Let's think step by step.
1. We have to find the number to multiply by 3.2 pages to get 12.8 pages.
2. «3.2 * 4 = 12.8», so we have to multiply 6.4 minutes by 4 to get how many minutes it takes to read 12.8 pages.
3. «6.4 * 4 = 25.6».
4. [Final solution] Thus, it takes her 25.6 minutes to read 12.8 pages.

- For step 1 of problems 7 and 8
  - Problem: Sarah is planning a road trip. It takes her 2.5 hours to drive 120 miles. If she wants to drive 240 miles, how long will it take her to reach her destination? Answer in decimal form.
  - Solution 1: Let's think step by step.
1. First, we have to find the speed Sarah drives.
2. The speed is «120 / 2.5 = 48» 48 miles per hour.
3. To calculate how much it takes for her to drive 240 miles, we should divide 240 miles by the speed. «240 / 48 = 5.0».
4. [Final solution] Thus, it takes her 5 hours to drive 240 miles.

  - Solution 2: Let's think step by step.
1. We have to find the number to multiply by 120 miles to get 240 miles.
2. «120 * 2 = 240», so we have to multiply 2.5 hours by 2 to get how many hours it takes to drive 240 miles.
3. «2.5 * 2 = 5».
4. [Final solution] Thus, it takes her 5 hours to drive 180 miles.

- For problem 9
  - Problem: Maya and Lily are having a friendly competition to see who bakes cookies faster. Maya already has a function for how long it takes her to bake cookies. Her function is y=3x+5, where x is the number of cookies and y is the minutes it takes her to bake x cookies. Lily takes 12 minutes to bake 4 cookies, and it takes her 2 minutes to set up. What is the slope of the function for Lily?
  - Solution 1: Let's think step by step.
1. We have to find an equation y = mx + b for Lily.
2. Since she takes 12 minutes to bake 4 cookies with 2 minutes to set up, the function will be «12 = 4m + 2».
3. If we isolate m, it will be «4m = 10».
4. Solving for m gives us «m = 10 / 4 = 2.5».
5. [Final solution] Thus, the slope of the function for Lily is 2.5.

  - Solution 2: Let's think step by step.
1. First, we have to find the time it takes Lily to bake 4 cookies.
2. The time is «12 - 2 = 10» 10 minutes.
3. To calculate the slope, we need to divide the time by the number of cookies.
4. The slope is «10 / 4 = 2.5».
5. [Final solution] Thus, the slope of the function for Lily is 2.5.

- For problem 9

– Problem: Carlos and Jane are trying to see who can paint a fence faster. Carlos has a function for how long it takes him to paint the fence. His function is y=5x+10, where x is the number of fence sections and y is the time it takes in minutes to paint x sections. Jane can paint 3 sections of fence in 20 minutes, and it takes her 5 minutes to set up. What is the slope of Jane's function?
– Solution 1: Let's think step by step.
1. We have to find an equation y = mx + b for Jane.
2. Since she takes 20 minutes to paint 3 sections with 5 minutes to set up, the function will be «20 = 3m + 5».
3. If we isolate m, it will be «3m = 15».
4. Solving for m gives us «m = 15 / 3 = 5».
5. [Final solution] Thus, the slope of Jane's function is 5.
– Solution 2: Let's think step by step.
1. First, we have to find the time it takes Jane to paint 3 sections.
2. The time is «20 - 5 = 15» 15 minutes.
3. To calculate the slope, we need to divide the time by the number of sections.
4. The slope is «15 / 3 = 5».
5. [Final solution] Thus, the slope of Jane's function is 5.

Solution 1 mimics the most popular student solutions. Solution 2 mimics the second most popular student solutions, if they exist, or are inspired by solutions to other problems. When we gave the same approach to LLMs, we always picked solution 1 for both problems. When we gave the different approaches to LLMs, we randomly picked solution 1 for one problem and solution 2 for the other.

C Results for Qwen2.5-Math-7B-Instruct

Table 18 shows the results for Qwen2.5-Math-7B-Instruct. We excluded the multi-turn reasoning, multi-agent, and multiple candidate methods from the results because the model did not follow those instructions. We also reduced the temperature from 1 to 0.7 because $t = 1$ had significant hallucinations.

			# slt.↑	Cov. all↑	Cov. maj↑	TVD↓	OOD↓
All problems							
Temperature sampling	$t = 0.3$		1.4	.58	**.81**	.38	.12
	$t = 0.7$		1.4	.56	.73	.43	.11
ICL	Random		1.3	**.62**	**.81**	.28	.11
	Related	Same	1.2	.54	.65	**.23**	**.03**
		Different	**1.6**	.60	**.81**	.48	.27
Multiple student solutions only							
Temperature sampling	$t = 0.3$		1.2	.40	**.72**	.42	.04
	$t = 0.7$		1.2	.36	.61	.50	.06
ICL	Random		1.2	**.44**	**.72**	.27	.03
	Related	Same	1.0	.33	**.72**	**.26**	**.02**
		Different	**1.6**	.42	**.72**	.43	.21

Table 18: Number of distinct LLM solutions (# slt.), coverage (Cov.) of all and majority (maj) student solutions, Total Variation Distance (TVD), and percent of LLM solutions outside student ones (OOD) for Qwen2.5-Math-7B-Instruct. Best in bold.

D Accuracy of LLM solutions

Table 19 shows the accuracy of LLM solutions for all problems. The accuracy of Qwen-2.5-7B-Instruct in the multi-agent setting is considerably low because, after a few turns, the model started to ignore the original instruction and kept generating summaries without proposing new solutions. These turns were labeled as incorrect.

E Example calculation of evaluation metrics

We show how to calculate our evaluation metrics in Section 3.2, using Problem 7 Step 1 in Table 13.

In this problem, asking GPT4o-mini for multiple solutions in one turn gave seven solutions in total. The distributions of students and GPT4o-mini are in Table 20. There are five non-zero rows for GPT4o-mini, so the number of distinct solutions is five. There are three student solutions (the top three rows), and, of those, GPT4o-mini has two non-zero rows (minutes per cup and ratio between cups). Therefore, the coverage of all solutions is $\frac{2}{3}$. When computing the coverage of majority solutions, we go through the rows from top to bottom until the sum of the proportions of student solutions exceeds 0.5. In this example, punch:time has less than 0.5, so we also consider minutes per cup. Since the sum of the proportions of punch:time and minutes per cup is $0.476 + 0.286 > 0.5$, we compute the coverage of these two solutions. Since GPT4o-mini did not give the punch:time solution, the coverage of the majority solutions is $\frac{1}{2}$. TVD is defined as half of the sum of the absolute differences between the students and GPT4o-mini

All problems			GPT4o-mini	Qwen2.5-7B-Instruct	Qwen2.5-Math-7B-Instruct	Gemma 3 12B
			Accuracy	Accuracy	Accuracy	Accuracy
Temperature sampling	$t = 0.3$		1.0	1.0	1.0	1.0
	$t = 1$		1.0	1.0	.92*	1.0
Multi-turn reasoning	ToT		.85	.85	—	.38
	Denial		1.0	1.0	—	1.0
	Paraphrase	Simple	1.0	.92	—	.92
		Replace	1.0	1.0	—	1.0
		Translate	1.0	1.0	—	1.0
Multi-agent	—	—	1.0	.23	—	1.0
Multiple candidates	Solution		1.0	.85	—	.92
	Approach		.85	1.0	—	.85
	Decompose		1.0	.69	—	.69
ICL	Random		1.0	1.0	.77	1.0
	Related	Same	1.0	1.0	.46	1.0
		Different	.85	1.0	.62	1.0

Table 19: Accuracy of LLM solutions for each method. *We used $t = 0.7$ for Qwen2.5-Math-7B-Instruct.

Annotation	Students	GPT4o-mini
punch:time	.476	0.0
Minutes per cup	.286	.143
Ratios between cups	.238	.429
Cups per minute	0.0	.143
Convert to seconds	0.0	.143
Cross multiply	0.0	0.0
Graph	0.0	.143

Table 20: Student and GPT-4o mini distributions for Problem 7 Step 1. The proportions of the solutions may not sum up to 1 due to rounding.

columns. Thus,

$$
\begin{aligned}
TVD = \frac{1}{2}(&|0.476 - 0| + |0.286 - 0.143| \\
&+ |0.283 - 429| + |0 - 0.143| \\
&+ |0 - 0.143| + |0 - 0| + |0 - 0.143|) \\
&= 0.597
\end{aligned}
$$

OOD is the sum of the proportions of GPT4o-mini solutions that students did not generate. In this example, the solutions not generated by students are cups per minute, convert to seconds, cross multiply, and graph. Therefore, $OOD = 0.143 + 0.143 + 0 + 0.143 = 0.429$.

F Results for problems with multiple student solutions

Tables 21-23 are the extensions of Tables 3-5 to the analysis only using the problems with multiple student solutions. In Table 21, most of the best methods for all problems are still the best in the problems with multiple student solutions, if not, remain the top 3. In Tables 22 and 23, the statistical significance of correlations is almost the same.

G An example of repetitive summarization of a conversation by Qwen2.5-7B-Instruct

Below is an example of Qwen2.5-7B-Instruct, where it kept summarizing a dialogue history in Problem 4 Step 1 in Table 9:

Turn 1: Sure! Here's my solution:

To determine how much of her battery Emma uses in 1 hour, we start by finding out how much battery she uses per hour based on the given rate.

Emma uses $\frac{1}{20}$ of her battery every $\frac{2}{3}$ of an hour. We need to find out how much she uses in 1 hour.

First, we set up the proportion:

$$\text{Battery used per hour} = \left(\frac{\frac{1}{20}}{\frac{2}{3}} \right)$$

To divide by a fraction, we multiply by its reciprocal:

$$\frac{1}{20} \div \frac{2}{3} = \frac{1}{20} \times \frac{3}{2} = \frac{3}{40}$$

So, Emma uses $\frac{3}{40}$ of her battery in 1 hour. Now, let's see if we can come up with a different solution that meets the criteria for being distinct.

Turn 2: Certainly! Let's explore a different approach to solve the problem.

			GPT4o-mini					Qwen2.5-7B-Instruct					Gemma 3 12B				
			# slt.↑	Cov. all↑	Cov. maj↑	TVD↓	OOD↓	# slt.↑	Cov. all↑	Cov. maj↑	TVD↓	OOD↓	# slt.↑	Cov. all↑	Cov. maj↑	TVD↓	OOD↓
All problems																	
Temperature sampling	$t = 0.3$		1.4	.52	.58	.54	.26	1.6	.60	.81	.44	**.14**	1.0	.34	.46	.47	.22
	$t = 1$		2.0	.60	.77	.50	.22	1.8	.55	.77	.53	.25	1.2	.40	.58	.46	.23
Multi-turn reasoning	ToT		1.2	.32	.38	.63	.23	0.8	.33	.50	.45	.19	0.9	.32	.42	**.20**	**.03**
	Denial		2.1	.61	.73	.63	.35	1.8	.61	.81	.47	.21	1.2	.45	.65	.46	.21
	Paraphrase	Simple	1.8	.42	.62	.53	.28	1.3	.51	.73	.45	.23	1.4	.39	.50	.47	.15
		Replace	1.6	.39	.42	.57	.35	1.5	.45	.69	.45	.24	1.9	.49	.62	.64	.43
		Translate	2.2	.60	.77	.56	.28	2.0	**.66**	**.85**	.51	.23	1.8	.52	.69	.47	.21
Multi-agent	—	—	**3.8**	.70	.81	.60	.43	**2.1**	.53	.81	.67	.38	**4.1**	**.72**	**.88**	.58	.39
Multiple candidates	Solution		3.5	**.73**	**.92**	.54	.42	1.4	.34	.50	.50	.40	**4.1**	.66	.81	.53	.40
	Approach		1.8	.60	.81	.56	.33	1.3	.40	.62	.50	.26	1.8	.43	.54	.55	.25
	Decompose		1.2	.51	.58	**.40**	**.09**	0.4	.14	.19	**.36**	.19	1.4	.45	.62	.38	.22
ICL	Random		1.5	.52	.65	.50	.17	1.7	.61	.81	.45	.18	1.5	.49	.77	.47	.20
	Related	Same	1.5	.57	.62	.45	.15	1.5	.46	.65	.46	.21	1.3	.51	.77	.46	.15
		Different	1.5	.40	.46	.53	.24	1.9	.61	**.85**	.46	.22	1.1	.40	.62	.46	.15
Multiple student solutions only																	
Temperature sampling	$t = 0.3$		1.2	.31	.39	.58	.18	1.6	.42	.72	.49	.07	1.0	.27	.44	.46	.13
	$t = 1$		1.8	.42	.67	.55	.17	1.9	.47	.78	.51	.13	1.3	.36	.61	.44	.14
Multi-turn reasoning	ToT		1.2	.35	.44	.63	.22	0.7	.25	.50	.43	.11	1.1	.24	.39	**.28**	.04
	Denial		2.0	.44	.61	.70	.31	1.8	.44	.72	.48	.12	1.3	.42	.72	.44	.10
	Paraphrase	Simple	1.8	.38	.67	.55	.23	1.1	.29	.61	.47	.17	1.6	.34	.50	.46	.17
		Replace	1.3	.35	.39	.61	.18	1.6	.43	.78	**.38**	.07	1.9	.38	.56	.63	.34
		Translate	1.9	.42	.67	.56	.18	2.0	.51	.78	.50	.13	1.7	.34	.67	.45	.17
Multi-agent	—	—	**3.6**	.56	.72	.60	.41	**2.2**	.43	.83	.61	.38	4.0	.60	.83	.51	.31
Multiple candidates	Solution		3.4	**.61**	**.89**	.54	.40	1.7	.39	.61	.47	.32	**4.4**	**.62**	.83	.55	.36
	Approach		1.8	.42	.72	.61	.30	1.3	.36	.67	.46	.11	1.8	.40	.56	.50	.13
	Decompose		1.1	.40	.50	**.41**	**.04**	0.6	.20	.28	.52	.28	1.6	.42	.67	.47	.24
ICL	Random		1.4	.42	.61	.54	.09	1.6	.44	.72	.43	.07	1.7	.49	**.89**	.43	.06
	Related	Same	1.3	.38	.44	.52	.10	1.6	.44	.72	**.38**	**.03**	1.2	.40	.78	.43	.01
		Different	1.3	.36	.44	.54	.12	2.1	**.54**	**.89**	.39	.06	1.0	.36	.67	.41	**0.0**

Table 21: The number of distinct LLM solutions (# slt.), LLMs' coverage (Cov.) of all and majority (maj) student solutions, Total Variation Distance (TVD), and percentage of LLM solutions outside student ones (OOD). Best in bold, top-3 with underline. The top half (all problems) is the same as Table 3.

Alternative Solution
Instead of directly using the proportion method, we can break down the time into smaller intervals and sum the battery usage over these intervals.
1. **Determine the number of $\frac{2}{3}$-hour intervals in 1 hour:**

$$\text{Number of intervals} = \frac{1 \text{ hour}}{\frac{2}{3} \text{ hours}} = \frac{1}{\frac{2}{3}} = \frac{3}{2} = 1.5$$

This means there are 1.5 intervals of $\frac{2}{3}$ hours in 1 hour.
2. **Calculate the total battery usage:** Since Emma uses $\frac{1}{20}$ of her battery in each $\frac{2}{3}$ of an hour, in 1.5 intervals, she will use:

$$\text{Total battery used} = 1.5 \times \frac{1}{20} = \frac{3}{2} \times \frac{1}{20} = \frac{3}{40}$$

Thus, Emma uses $\frac{3}{40}$ of her battery in 1 hour.
Critique and Distinctness
- **Methodology:** - Your original solution used a direct proportional relationship. - My alternative solution broke down the time into smaller intervals and summed the usage over these intervals.
- **Intermediate Steps:** - Your solution involved a single division step to find the hourly usage. - My solution involved calculating the number of intervals and then multiplying by the battery usage per interval.
- **Assumptions:** - Both solutions assume the same rate of battery usage.
- **Generalization:** - Your solution generalizes well to any similar problem involving rates over fractional time periods. - My solution also generalizes similarly but might be seen as more detailed in its breakdown.
- **Simplicity:** - Both solutions are relatively simple, but your solution might be considered slightly more straightforward due to its direct application of proportions. This alternative solution meets the criteria

	GPT4o-mini & Qwen2.5-7B-Instruct	GPT4o-mini & Gemma 3 12B	Qwen2.5-7B-Instruct & Gemma 3 12B
All problems			
# slt.	.46 (.101)	.92 (<.001)*	.23 (.424)
Cov. all	.10 (.721)	.72 (.003)*	-.01 (.966)
Cov. maj	.11 (.713)	.54 (.045)*	.13 (.655)
TVD	.46 (.096)	.02 (.948)	.42 (.131)
OOD	.72 (.003)*	.63 (.015)*	.69 (.007)*
Multiple student solutions only			
# slt.	.47 (.089)	.93 (<.001)*	.29 (.312)
Cov. all	.07 (.818)	.92 (<.001)*	.19 (.526)
Cov. maj	.03 (.931)	.46 (.100)	.23 (.431)
TVD	-.04 (.886)	-.09 (.747)	.06 (.845)
OOD	.47 (.091)	.58 (.030)*	.71 (.005)*

Table 22: Pearson's correlations between two LLMs for each metric for all problems (top) and problems with multiple student solutions (bottom). $p < .05$ is marked with *. The top half is the same as Table 4.

Pearson's r (p-value)		All	Multiple
GPT4o-mini	Cov. all	.78 (.001)*	.93 (<.001)*
	Cov. maj	.76 (.002)*	.80 (<.001)*
	TVD	.32 (.261)	.18 (.534)
	OOD	.83 (<.001)*	.85 (<.001)*
Qwen2.5-7B-Instruct	Cov. all	.89 (<.001)*	.93 (<.001)*
	Cov. maj	.92 (<.001)*	.91 (<.001)*
	TVD	.66 (.011)*	.17 (.565)
	OOD	.25 (.398)	.02 (.959)
Qwen2.5-Math-7B-Instruct	Cov. all	.53 (.354)	.63 (.259)
	Cov. maj	-.05 (934)	.06 (.920)
	TVD	.93 (.020)*	.51 (.380)
	OOD	.97 (.006)*	.92 (.025)*
Gemma 3 12B	Cov. all	.92 (<.001)*	.86 (<.001)*
	Cov. maj	.71 (.004)*	.54 (.047)*
	TVD	.53 (.052)	.53 (.051)
	OOD	.77 (.001)*	.76 (.002)*

Table 23: Pearson's correlations with the number of distinct LLM solutions for all problems (left) and problems with multiple student solutions (right). $p < .05$ is marked with *. The left half is the same as Table 5.

for being distinct from your original solution.

Turn 3: Thank you for providing the alternative solution. Let's critique it further and ensure it meets all the criteria for being distinct.
Critique of the Alternative Solution
1. **Methodology:**
- Your original solution used a direct proportional relationship to find the battery usage per hour.
- The alternative solution broke down the time into smaller intervals and summed the battery usage over these intervals.
2. **Intermediate Steps:**
- Your solution involved a single division step to find the hourly usage.
- The alternative solution involved calculating the number of intervals (1.5) and then multiplying by the battery usage per interval ($\frac{1}{20}$).
3. **Assumptions:**
- Both solutions assume the same rate of battery usage, which is $\frac{1}{20}$ of the battery every $\frac{2}{3}$ of an hour.
4. **Generalization:**
- Both solutions generalize well to similar problems involving rates over fractional time periods.
- However, the alternative solution might be seen as more detailed in its breakdown, which could make it easier to understand the process step-by-step.
5. **Simplicity:**
- Both solutions are relatively simple, but your original solution might be considered slightly more straightforward due to its direct application of proportions.
New Solution
Let's propose a new solution that is fundamentally different from both previous ones. We'll use a geometric interpretation to solve the problem.
Geometric Interpretation Solution
1. **Visualize the Problem:**
- Imagine a rectangle representing 1 hour of battery usage.
- Divide this rectangle into three equal parts, each representing $\frac{1}{3}$ of an hour.
- Each part is further divided into two equal segments, each representing $\frac{1}{6}$ of an hour.
2. **Determine Battery Usage per Segment:**
- Since $\frac{2}{3}$ of an hour corresponds to $\frac{1}{20}$ of the battery, we need to find how much battery is used in $\frac{1}{6}$ of an hour.
- $\frac{2}{3}$ of an hour is equivalent to 2 segments of $\frac{1}{6}$ of an hour.
- Therefore, $\frac{1}{20}$ of the battery is used in 2 segments of $\frac{1}{6}$ of an hour.
- So, in one segment of $\frac{1}{6}$ of an hour, the

battery usage is:

$$\frac{1}{20} \div 2 = \frac{1}{40}$$

3. **Calculate Total Battery Usage in 1 Hour:**
- There are 6 segments in 1 hour (since $3 \times 2 = 6$).
- Therefore, the total battery usage in 1 hour is:

$$6 \times \frac{1}{40} = \frac{6}{40} = \frac{3}{20}$$

Critique of the New Solution
- **Methodology:**
- The original solution used a direct proportional relationship.
- The alternative solution broke down the time into smaller intervals and summed the battery usage over these intervals.
- The new solution uses a geometric interpretation to visualize and break down the problem.
- **Intermediate Steps:**
- The original solution involved a single division step.
- The alternative solution involved calculating the number of intervals and then multiplying by the battery usage per interval.
- The new solution involves visualizing the problem and breaking it down into segments.
- **Assumptions:**
- All solutions assume the same rate of battery usage.
- **Generalization:**
- All solutions generalize well to similar problems involving rates over fractional time periods.
- The new solution might be seen as more intuitive for some students due to its visual representation.
- **Simplicity:**
- All solutions are relatively simple, but the new solution might be considered more intuitive and easier to visualize.
This new solution meets the criteria for being distinct from both previous solutions.

Turn 4: Your new solution using a geometric interpretation is indeed distinct and provides a unique perspective on the problem. Let's summarize and critique it further to ensure it meets all the criteria for being distinct.
Summary of the New Solution
1. **Geometric Interpretation:**
- Visualize 1 hour as a rectangle.
- Divide this rectangle into three equal parts, each representing $\frac{1}{3}$ of an hour.
- Each $\frac{1}{3}$ of an hour is further divided into two equal segments, each representing $\frac{1}{6}$ of an hour.
- Since $\frac{2}{3}$ of an hour corresponds to $\frac{1}{20}$ of the battery, we need to find how much battery is used in $\frac{1}{6}$ of an hour.
- $\frac{2}{3}$ of an hour is equivalent to 2 segments of $\frac{1}{6}$ of an hour.
- Therefore, $\frac{1}{20}$ of the battery is used in 2 segments of $\frac{1}{6}$ of an hour.
- So, in one segment of $\frac{1}{6}$ of an hour, the battery usage is:

$$\frac{1}{20} \div 2 = \frac{1}{40}$$

- There are 6 segments in 1 hour (since $3 \times 2 = 6$).
- Therefore, the total battery usage in 1 hour is:

$$6 \times \frac{1}{40} = \frac{6}{40} = \frac{3}{20}$$

Critique of the New Solution
1. **Methodology:**
- **Original Solution:** Used a direct proportional relationship.
- **Alternative Solution:** Broke down the time into smaller intervals and summed the battery usage over these intervals.
- **New Solution:** Used a geometric interpretation to visualize and break down the problem.
2. **Intermediate Steps:**
- **Original Solution:** Involved a single division step.
- **Alternative Solution:** Involved calculating the number of intervals and then multiplying by the battery usage per interval.
- **New Solution:** Involved visualizing the problem and breaking it down into segments.
3. **Assumptions:**

- All solutions assume the same rate of battery usage, which is $\frac{1}{20}$ of the battery every $\frac{2}{3}$ of an hour.
4. **Generalization:**
- All solutions generalize well to similar problems involving rates over fractional time periods.
- The new solution might be seen as more intuitive for some students due to its visual representation.
5. **Simplicity:**
- All solutions are relatively simple, but the new solution might be considered more intuitive and easier to visualize.
Conclusion
The new solution using a geometric interpretation is fundamentally different from both the original and the alternative solutions because it relies on a visual and spatial understanding rather than algebraic manipulation or interval summation. It provides a distinct approach that can offer clarity and intuition to students who might find the algebraic or interval-based methods less intuitive.
Therefore, the new solution meets all the criteria for being distinct and offers a valuable alternative method for solving the problem.

Turn 5: Thank you for the detailed summary and critique. Let's formalize the conclusion and ensure that the new solution is clearly distinct from the previous ones.
Final Summary and Conclusion
Original Solution
- **Methodology:** Direct proportional relationship.
- **Intermediate Steps:** Single division step.
- **Assumptions:** Same rate of battery usage.
- **Generalization:** Generalizes well to similar problems.
- **Simplicity:** Slightly more straightforward due to direct application of proportions.
Alternative Solution
- **Methodology:** Breakdown into smaller intervals and summation.
- **Intermediate Steps:** Calculate number of intervals and multiply by battery usage per interval.
- **Assumptions:** Same rate of battery usage.
- **Generalization:** Generalizes well to similar problems.
- **Simplicity:** More detailed but still straightforward.
New Solution (Geometric Interpretation)
- **Methodology:** Geometric visualization and breakdown.
- **Intermediate Steps:** Visualize 1 hour as a rectangle, divide into segments, and calculate battery usage per segment.
- **Assumptions:** Same rate of battery usage.
- **Generalization:** Generalizes well to similar problems.
- **Simplicity:** More intuitive and easier to visualize.
Distinctness Criteria Met
1. **Fundamentally Different Methodology:**
- The original solution uses algebraic manipulation.
- The alternative solution uses interval summation.
- The new solution uses geometric visualization.
2. **Significantly Different Intermediate Steps:**
- The original solution involves a single division.
- The alternative solution involves calculating intervals and multiplying.
- The new solution involves visualizing and breaking down the problem into segments.
3. **Different Assumptions:**
- All solutions assume the same rate of battery usage, so this criterion is not applicable here.
4. **Generalization:**
- All solutions generalize well to similar problems, so this criterion is not applicable here.
5. **Simplicity:**
- While all solutions are simple, the new solution is more intuitive and easier to visualize, making it distinct in terms of simplicity.
Conclusion The new solution using a geometric interpretation is fundamentally

different from both the original and the alternative solutions. It provides a distinct approach that leverages visual and spatial understanding, making it a valuable alternative method for solving the problem. This solution meets all the criteria for being distinct and offers a unique perspective that can enhance understanding for students.

Thus, the new solution is indeed distinct and valuable.

Turn 6: ### Final Summary and Conclusion
<the same final summary as Turn 5>

Qwen2.5-7B-Insturct stopped proposing new solutions in Turn 4 and generated the final summary in Turn 5. After Turn 6, it kept generating the same final summary.

H An example of flawed reasoning in multi-agent debate

Below is an example where the multi-agent debate method gave the correct final answer with flawed reasoning:

1. We know that Emma uses $\frac{2}{15}$ of the battery in half an hour.

2. Let $B(n)$ represent the battery level after n half-hour increments: At $n = 0$: $B(0) = 1$ (full battery). At $n = 1$: $B(1) = B(0) - \frac{2}{15} = 1 - \frac{2}{15} = \frac{13}{15}$. ... At $n = 7$: $B(7) = B(6) - \frac{2}{15} = \frac{3}{15} - \frac{2}{15} = \frac{1}{15}$. At $n = 8$: $B(8) = B(7) - \frac{2}{15} = \frac{1}{15} - \frac{2}{15} = -\frac{1}{15}$.

3. ... After 7 increments (3.5 hours), the battery is at $\frac{1}{15}$. After 8 increments (4 hours), the battery is depleted.

4. Therefore, the total time it takes to use the entire battery is: $4\,\text{hours} = \frac{15}{4}$ hours.

The last step does not follow Step 3 because the total time must be somewhere between 3.5 and 4 hours, and is arithmetically incorrect because 4 is not $\frac{15}{4}$. Still, this solution gives the correct answer, which is $\frac{15}{4}$ because LLMs know it from previous turns.