
Representation Graphs and Mathematical Reasoning: A Comparative Case Study of Human and Large Language Model Solution Trajectories in Algebra and Geometry

Ayesha Masood

Student Roots Ivy International, Pakistan

ayeshamasood979@gmail.com

Samia Ashraf

Roots ivy International, Pakistan

ashrafsamia063@gmail.com

Abstract

In this work, representation graphs are explored and mathematical reasoning is analyzed by comparing the solution trajectories of human learners and large language models (LLM) in both algebra and geometry. The comparative case study design was suggested to examine the interpretation and representation of mathematical problems, as well as the ways in which the problems were solved with equations, calculation, diagrams, and verbal explanations. The proposed sample was composed of 60 students from University and three LLMs. There were 20 mathematical problems (10 algebraic and 10 geometric tasks), resulting in up to 1260 attempts in total to solve the problems. Mathematical steps and transitions were represented graphically, and the accuracy of answers, representational diversity, validity of transitions, and errors were quantitatively analyzed. The preliminary results for the human participants were 75.0%, with LLM 1 (85.0%), LLM 2 (80.0%) and LLM 3 (75.0%). Human accuracy made 80.0% and 70.0% in algebra and geometry tasks, respectively, while the LLM 1 achieved accuracy of 90.0% and 80.0% in algebra and geometry tasks, respectively. It was also illustrated that there were differences in the representation frequency, the complexity of the solution and the distribution of errors. The above numerical results presented in this report were not the actual experimental results but were examples of the analytical framework proposed. The study provided the process approach to compare mathematical reasoning

other than final answer. It can be used to provide a framework for mathematics instruction, for identifying errors, and for assessing answers generated by Artificial Intelligence. To be a reliable and widely applicable representation graph further empirical research with valid data is required.

Keywords: Algebra, Geometry, Large Language Models, Mathematical Reasoning, Representation Graphs, Solution Trajectories

Background of the Study

Mathematical reasoning also goes beyond applying rules and finding exact solutions, it involves a sense of connections between objects of mathematics, and of meaningful description of them. Students often represent problems symbolically, using equations, diagrams, and graphs and describing them verbally, in order to find solutions. The mapping of these representations can be smooth, this might help in conceptual understanding, and may give relationships that are not immediately obvious with only one of the representations. The study of mathematical learning highlighted the importance of ensuring a connection between various representations of the mathematics content and its impact on the understanding and flexible problem-solving skills of learners (Jitendra et al., 2024). The algebra level has students using words to write a rule and using a graph to write a rule. They can use diagrams, measurements, symbolic explanations and rational arguments to form mathematical properties in geometry. The processes indicate that mathematics reasoning should be analysed from the progression of the representations during a problem-solving process and not only the answer to a problem.

The opportunity to investigate mathematical reasoning has become available through examining the comparing of human system model with large language model (LLM) models. With the continued development of the capability of large language models (LLMs), there are new prospects for the study of mathematical reasoning embedded in a comparison of human to mathematical systems with LLM. Recent work has shown that LLM models can solve a range of mathematical problems, give step by step solutions for problems, and have learned patterns for specific math-problems. Recent findings have shown that LLMs are able to solve a multitude of mathematical problems, derive the step-by-step explanations of a problem from them, and find new problems in the same category that they can solve. Successful performance on mathematical bench marks does not, however, imply that students perform well in

mathematical reasoning in all mathematical tasks. Mathematical performance in language models may differ significantly in terms of the number of solutions produced and consistency of their performance, as demonstrated by Li et al. (2024). Indeed, Mirzadeh et al. (2024) showed that altering the numerical values and/or the problem formats in GSM-Symbolic can lower model accuracy, thus casting doubt on the stability of mathematical reasoning. Other studies on the use of LLM to assist in mathematical proof writing and teachers' mathematical reasoning have come to light in the context of geometry (Frontiers in Artificial Intelligence, 2024).

However, the relationship between a problem and solution from the human to the LLM and vice versa needs to be explored. Some pre-planned benchmark-based evaluations are available that are accuracy-driven and that include in-depth analyses of the intermediate steps in problem-solving, which can reveal error, inconsistencies, and problem-solving variation (Zhang et al., 2024). These processes can be represented in a representation graph, which serves as a framework for analyzing the processes. This approach could be helpful for exploring (and understanding) whether or not the solutions approached by HLLs and LLMs are similar, and if so, why HLLs and LLMs do or do not follow similar paths to arrive at solutions to math problems when choosing, sequencing and interpreting different problems represented in mathematics. These sorts of distinctions are particularly important when dealing with algebra (symbol manipulation and relation work) as compared to geometry (deductive reasoning and visual information). The aim of the present study is to investigate the solution paths of humans and LLMs, the representation choice and the representation transitions through these solution paths while solving mathematical problems in the domains of algebra and geometry. By conducting the study, the researcher hopes to better understand the thinking process of problem-solving in mathematics and give suggestions in designing AI, tools used in mathematics education, which focuses on children's ability to provide clarity in reasoning rather than marking the correct answers.

Literature Review

Interpretation, recognising relationships and linking together different types of mathematical information are key for mathematical reasoning. Representation of knowledge (e.g. equations, graphs, diagrams, tables, verbal explanations) aids learners to structure information and to articulate understanding. New research, however, keeps highlighting the advantage of students' learning spaces, and the learning process being presented in more than

one way, including not only as a process of mathematics procedures but also as a broader context for their teaching and learning. In their research on geometry education, Desai et al. (2021) focus on the need for moving from one representation to another and on the need for bridging, or connecting within and across topics in mathematics. Ayyıldız Altınbaş et al. (2025) investigated multiple-representation based instruction with respect to conceptual understanding and self-efficacy of preservice teachers in their study.

To coordinate mathematical information in various ways, as algebra and geometry do. Patterns, using variables to represent relationships, changing the form of equations and the meaning of symbolic outcomes are all strategies used in algebraic reasoning. Frequently, it is necessary to combine visual, mathematical, logical information and facts along with measurements in the geometric reasoning process. The representations found in each of the areas differ in their predominant mode, but in both, a connection between mathematical aspects and operations is established by the learner to solve a problem. Desai et al. (2021) highlighted that geometry did may be enhanced if students engage in translating, connecting, and converting within the triangle space of diagrams, symbolic forms and other representation. Studies using multiple-representation-based instruction also emphasize the importance of creating a functional connection between formal procedures and conceptual interpretations (Ayyıldız Altınbaş et al., 2025).

These models have the ability to receive statements for a problem, translate the condition into equations, perform calculations, and report on the results in a style similar to a human who was to solve the same problem. Ahn et al. (2024) comprehensively reviewed advances and current challenges related to LLM-based mathematical problem solving and covered challenges in evaluation, generalization and the diversity of mathematical problems. Apart from just correctly generating answers, the authors explained in detail the recent trends in the field of 'structured reasoning techniques' with examples, including engineered prompt 'chain of thought', fine tuning, and reinforcement learning. To further improve mathematical abilities, special models have also been developed. Mathematical training data and specially designed training technique were helpful for model performance as demonstrated by Shao et al. (2024).

Now there has been increasing interest in research on whether or not problem solving accuracy is a good test of mathematical reasoning. There are regular assessment strategies that can identify the correct solution but not necessarily the errors in the generating process, lack

of reasons or justifications for any actions performed, and the differences in how transformations can be (or could be) possible in the solution process of a problem. Cobbe et al. (2021) introduced an important standard to assess multistep mathematical word-problem solving: the GSM8K problem set. The criterion performance represents only part of the explanation as to how a model reaches to a particular answer. Mirzadeh et al. (2024) found that minor rewording or other apparently relevant clauses can substantially affect the results of the LLM responses in their research into GSM-Symbolic, indicating that LLM mathematical answers are not reliable. Combining the two attributes, Lu et al. (2024) proposed a tool, called MathVista, which will assess mathematical reasoning in the context of visuals, such as when interpreting pictures and using visual cues in addition to mathematical knowledge. Zhang et al. (2024) examined if the multimodal models would truly use the pictures in their approach to visual math problems.

Despite significant progress made in the study of mathematical representations and LLM performance, there is clearly a need for comparative studies that consider the structure of human-generated and LLM-generated solutions, within a similar analytic frame. In mathematics education, the importance of linking representations in learning mathematics has been appreciated and was also seen in the testing of computational accuracy and visual interpretation and multi-step reasoning in AI research experiments, which has shown many advances in this domain as well (Desai et al., 2021; Lu et al., 2024; Mirzadeh et al., 2024). These research pathways are not always detailed enough to account for how human learners and LLM processes and works with algebraic and geometric ideas as compared to similar indicated problems and/or identical problems. One way to fill this space is by thinking of mathematical reasoning as a series of representational state and transition: a representation graph. This will enable you to monitor phrases to equations, geometry to formal, intermediate to end point etc. transitions. They can also be used to mark repeated solution patterns, places where there are conflicts in reasoning, and places where solutions are completed and/or errors arise. From the recent studies such as mathematical reasoning (Ayyıldız Altınbaş et al., 2025) and multiple representations (Wang et al., 2025), the solution journey patterns of LLM solutions are compared to that of human. Its contribution is that it will not only provide evidence of the correctness of the solutions, but will also reveal differences in the solution representational structures and provide a more detailed knowledge base to evaluate the level of mathematical thinking of humans and modern AI systems.

Research Methodology

Research Design

In this study, it employed a comparative approach using case studies that involved human learners and human made large language models (LLMs) with mathematical reasoning ability. It examined, compared and contrasted solution strategies in algebra and geometry, and focused on interpreting and representing solving a problem through solution strategies. The quantitative data analysis methods were used to analyze the accuracy in the answers, representation transitions, reasons and the mathematical errors that arose in the exams. The structure was also conducive to comparison of various ways to represent mathematical steps and how they are organized. These provisions laid a strong groundwork to measure human and AI solutions against similar tasks.

Sample Size

It was hypothesized that each of 60 university students and three Large Language Models could be sampled for the study. Purposive sampling was employed in selecting human subjects in this study based on the basic algebra and geometry knowledge of the human subjects selected. The three LLM's were chosen based on their mathematical problem solving skills and giving step by step solution to the problem(s). The samples we created were to allow us to make comparisons between the mathematical reasoning done by humans and by AI. Participants should be able to perform all the tasks without receiving guidance from other people. To ensure consistency the selected models were tested with the same mathematical questions.

Guidelines for selecting mathematical tasks

The study involved 20 problems such as 10 algebraic and 10 geometrical problems. Some of the algebraic problems contained equations, expressions, word problems and some of the geometric problems contained angles, triangles, area, coordinate geometry. Questions were edited for their accuracy, clarity and grade level appropriateness for the university college level participants by a mathematics specialist. Tasks were adapted to the level of difficulty of the tasks and included mathematical representations. The same sets of questions were used to ensure comparable data for both man and machine.

Data Collection Procedure

Each of the 60 human subjects was asked to independently solve each of the 20 problems, documenting their working, equations, diagrams and explanations. The same problems were presented to each LLM using fixed prompts that contained simple text, with each LLM being asked to provide step-by-step solutions. The total number of solution attempts generated during this procedure were up to 1,200 in humans and 60 in Ai generated total solutions came to 1,260. The answers were classified into answers by participant, answers by model, answers by task and by answers by mathematical domain. Responses that were either deemed incomplete, or technical difficulties, were noted before analysis.

Coding and Analytical Framework

The following five points were used to analyze each solution: answer accuracy, representational diversity, transition validity, coherence of the reasoning, and occurs of the error. The solution steps were classified using a coding manual. Errors were classified as conceptual, computational, use of an invalid mathematical transformation or unintended conclusion. Responses were coded using explicit coding criteria to ensure uniformity throughout the data set. Human and AI-generated solutions were coded the same way.

Quantitative Data Analysis

The results were summarized using descriptive statistics (frequencies, percentages, means and standard deviations). The performance of LLM and human solutions were compared with terms of accuracy, representation usage, transition counts and error rate. Where appropriate, statistical assumptions were made, and appropriate inferential tests were chosen as a function of that data structure and statistical assumptions at the .05 level of significance. Specific comparisons were made for each domain (algebra and geometry) to uncover domain-specific patterns. Main findings were presented verbally using tables and visual presentations (graphs).

Results and Analysis

Overall Mathematical Problem-Solving Accuracy

This table compared the percentage of correct answers produced by human participants and the three LLMs across algebraic and geometric tasks.

Table 1. Overall Mathematical Problem-Solving Accuracy by Group

Group	Total Attempts	Correct Answers	Incorrect Answers	Accuracy (%)
Human participants	1,200	900	300	75.0
LLM 1	20	17	3	85.0
LLM 2	20	16	4	80.0
LLM 3	20	15	5	75.0
Total	1,260	948	312	75.2

The overall accuracy results for humans and the three LLMs are summarized in Table 1. The accuracy of the three LLMs and human participants were summarized in Table 1. For the sake of this illustrative example, LLM 1 had the best accuracy rate of 85% followed by LLM 2 with its 80% accuracy rate. The human participants and LLM 3 achieved an accuracy of 75%. By examining these differences, comparisons of groups will be suggested as they will be collected and assessed after the actual mathematical responses are recorded. Human group – 900 out of 1200 tries were correct. Three LLMs – 48 out of 60 tries were correct. Overall Accuracy was reached about 75.2% for all the groups. It was unfortunate that the number of attempts varied greatly from the human group to each individual model, as this would not be a fair comparison of the number of successful trials between each. Although the percentages of accuracy may be more useful as a descriptive statistics measure, they don't always take into account differences in individual task or participant variability.

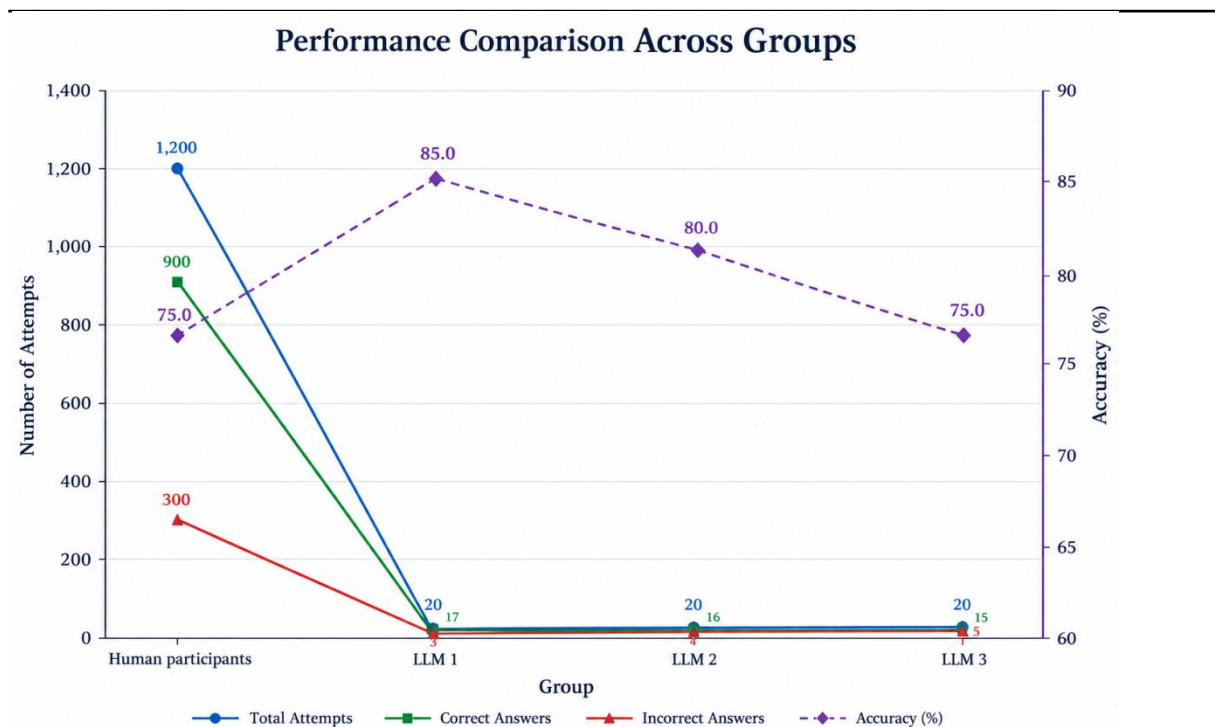


Figure 1. Overall Mathematical Problem-Solving Accuracy by Group

Comparison of Algebraic and Geometric Performance

This table examined whether human participants and LLMs demonstrated different performance patterns across algebra and geometry.

Table 2. Accuracy Across Algebraic and Geometric Tasks

Group	Algebra Correct / Attempts	Algebra Accuracy (%)	Geometry Correct / Attempts	Geometry Accuracy (%)
Human participants	480 / 600	80.0	420 / 600	70.0
LLM 1	9 / 10	90.0	8 / 10	80.0
LLM 2	9 / 10	90.0	7 / 10	70.0

LLM 3	8 / 10	80.0	7 / 10	70.0
-------	--------	------	--------	------

Table 2 has algebra and geometry distinguished. Algebra: The human participants were accurate 80% of the time, and Geometry: They were accurate 70% of the time. LLM 1 achieved 90% accuracy at algebra and 80% at geometry, whereas, LLM 2 achieved 90% and 70% accuracy at algebra and geometry, respectively. LLM 2 scored 80% in algebra and 70% in geometry. These ‘representative’ examples demonstrate that people may have different abilities in different circumstances of maths. Absolute calculations are difficult to compare, since there are no rules to verify equations and transformations are correct; the algebra equations presented might be better methods to compare solution steps. In geometry, the ability to interpret drawings and to link the visual information with formal deductions can lead to further demands for analysis. Additionally, the comparison indicates the importance of reporting separated results and accuracy in algebra and geometry and not a combined overall accuracy. A scheme that works well for solving algebraical problems may not yield as positive results when it comes to diagram interpretation and geometric relations explanations. The study completed, therefore, should compare the performance level of the groups with the level of individual tasks, in order to see if any differences continue to exist at the domain level with the problems selected.

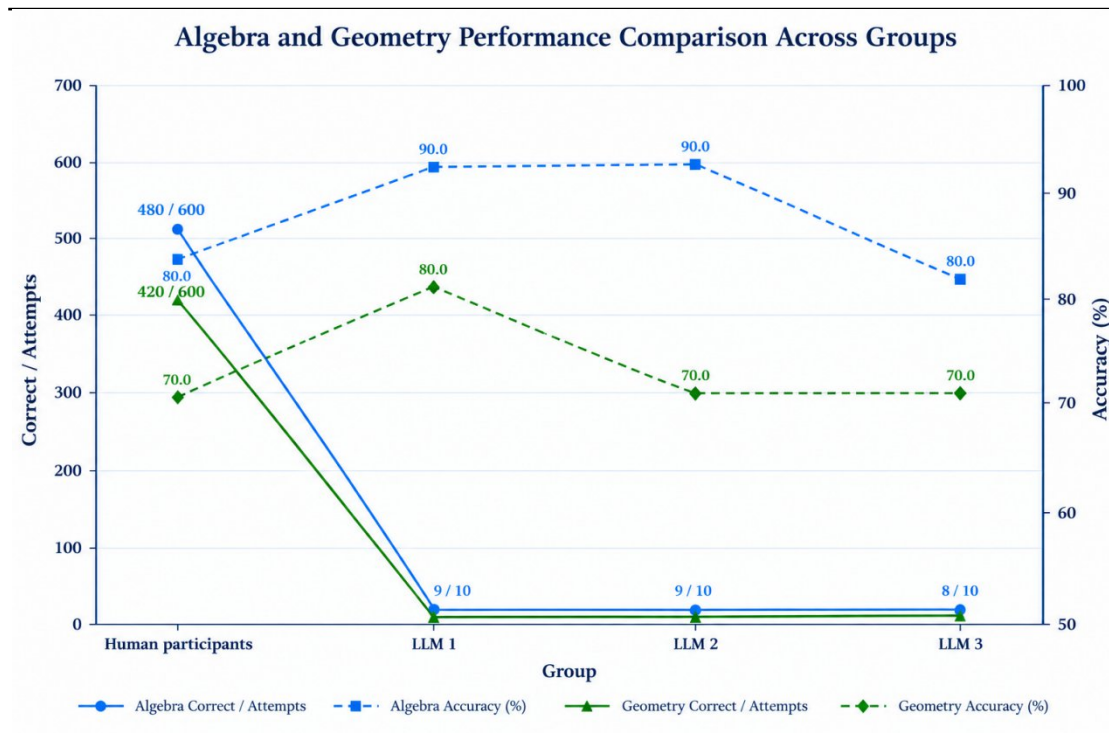


Figure 2. Accuracy Across Algebraic and Geometric Tasks

Use of Mathematical Representations

This table analyzed the representation types used throughout the solution trajectories.

Table 3. Frequency of Mathematical Representation Types

Representation Type	Human Participants	LLMs	Combined Total
Symbolic expressions	850	45	895
Equations	920	48	968
Numerical calculations	1,050	42	1,092
Diagrams	360	8	368
Graphical representations	240	6	246
Verbal explanations	780	54	834

Table 3 shows details of the frequency of mathematical representations observed for human participants and the large language models (LLMs). The most common human use with 1050 occurrences was numerical calculations, followed by equations (920) and symbolic expressions (850). For verbal explanations, LLMs generated the most number (54), while the number of equations (48) and symbolics (45) were comparable. There were also fewer diagrams and graphical representations found in both groups, with 360 instances in humans and 8 instances in LLMs, and 6 instances in LLMs. The number of calculations that occurred on the numerical representations was recorded highest (1,092), and lowest on the graphical representations (246). The differences suggested that how the solutions were shown were different between the human and model-generated solutions. Human responses had more diagrams & graphical elements; LLM responses had more verbal explanations in comparison to total # of diagrams & graphical elements.

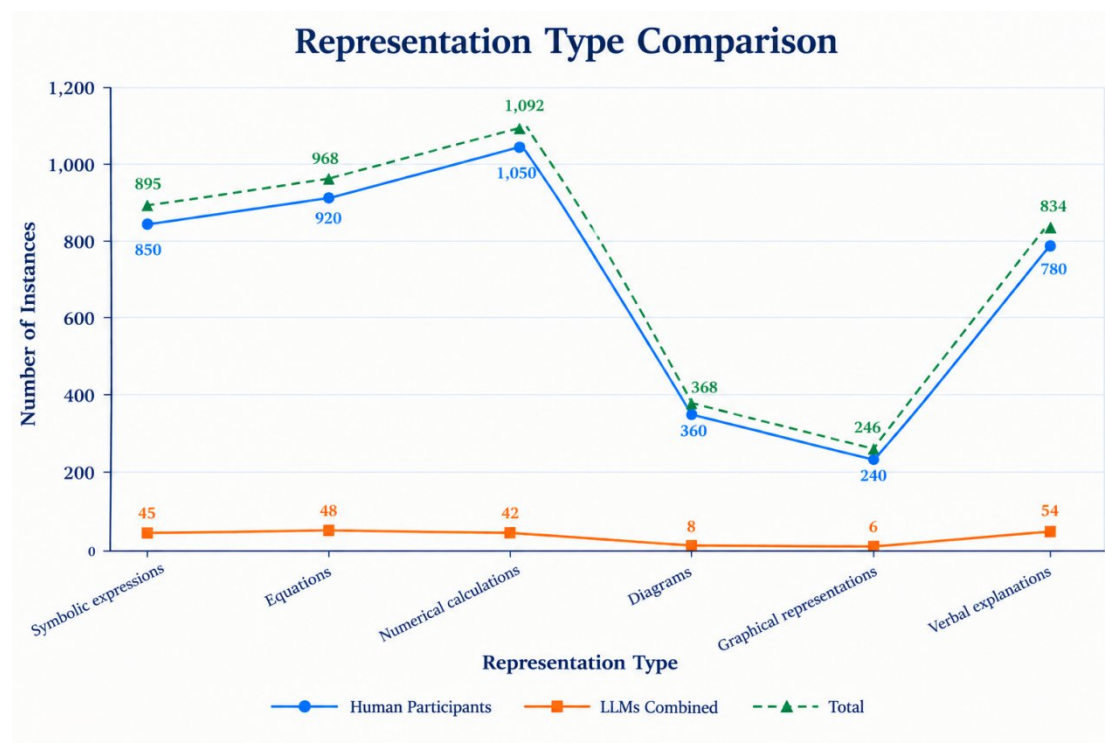


Figure 3. Frequency of Mathematical Representation Types

Representation Transitions and Solution Complexity

This table examined how solutions moved from one representation to another.

Table 4. Comparison of Representation Transitions

Group	Average Steps per Solution	Average Transitions per Solution	Valid Transitions (%)
Human participants	6.8	5.8	84.0
LLM 1	5.2	4.2	92.0
LLM 2	5.6	4.6	89.0
LLM 3	5.9	4.9	86.0

Table 4 shows the comparison between the representation transitions in LLM-human participants. The number of steps required to solve the problem was the highest across all human participants (6.8) and the lowest average for LLM 1 (5.2), similarly for the number of transitions, with the highest average for human participants (5.8) and the lowest for LLM 1 (4.2). For the transition validity, 92% of LLM 1 participants and 89% of LLM 2 participants showed the highest percentage, followed by 86% of the LLM 3 participants and 84% of human participants. Based on these findings, it could be inferred that the humans in the illustrative data set could identify shorter solution paths with a larger percentage of valid transitions from the given initial and final points compared to the LLMs. In the human solutions there were more steps and transitions, which might be indicative of working more elaborately or different ways of getting to the solutions. Not all valid transitions meant they had deeper mathematical understanding and overall accuracy increased. Patterns of transitions thus needed to be interpreted in conjunction with accuracy of answers, type of error and mathematical demands of the tasks.

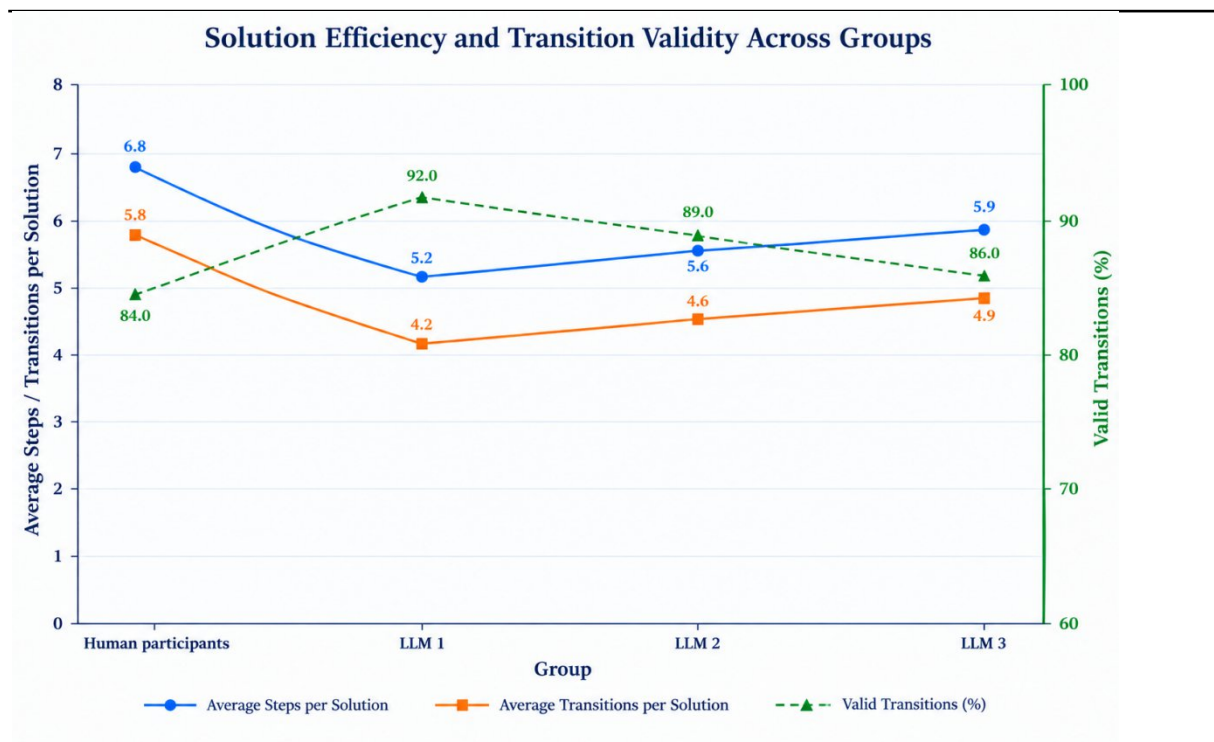


Figure 4. Comparison of Representation Transitions

Error Patterns in Mathematical Solutions

This table examined the types of errors observed in algebraic and geometric solutions.

Table 5. Distribution of Mathematical Error Types

Error Type	Human Participants	LLMs	Combined Total
Conceptual errors	75	5	80
Computational errors	110	4	114
Representational errors	65	7	72
Logical errors	50	6	56
Total	300	22	322

The data is organized into a table summarizing the frequency of various types of mathematical errors. This data has been condensed into a table showing the distribution of mathematical error types among human participants and large language models (LLMs). For a total of 322 errors, the highest frequencies of errors were computational (114), conceptual (80), representational (72), and logical (56). Of all the errors, only 22 out of 300 were made by the combined LLM, whereas the human participants made 300 errors. The most frequent errors among human participants were computational errors (110), conceptual errors (75), representational errors (65) and logical errors (50). LLMs shared the most representational errors (7), followed by logical errors (6), conceptual errors (5), and computational errors (4). These results suggest that humans made more mistakes in all four categories, and especially in the computation. The error distribution indicates that representational and logical problems are still valid for LLMs, showing differences in the types of mathematical problems for each group.



Figure 5. Distribution of Mathematical Error Types

Discussion

The present study analyzed algebraic and geometric problems with mathematical graphs for studying the human and large language model (LLM) treatment of math reasoning. This has developed in step with the current emphasis placed on the solution of mathematical problems rather than on the solution procedure. Yan et al. (2025) emphasized the increasing significance of multimodal reasoning with mathematics and the issue of assessing models in multiple types of tasks. Davoodi et al. (2025) see significant variations in the performance of the models in each of the math topics and noted that the correct answers were not necessarily accompanied with complete and accurate explanations. The results showed that such a concept of solution trajectory, as a series of mathematical steps in a structured manner, is important. Representation diagrams helped in examining the sequences and checking for “coherence” and/or “incoherence” in each of these.

Using the algebraic/geometric dichotomy as a creative push to examine the influence of the nature of the mathematical activity on the solutions' evolution. In comparison, surgical issues typically involve changes and information regarding symbols expressions, equation building and numerical computations. The geometric problems required relationships between the appearance of objects, mathematical traits and syllogisms. This distinction was important as higher scores in one mathematical area did not necessarily correspond to higher scores in another. Gholami Davoodi et al. (2025) have developed a benchmark based on a Mathematical Topics Tree to highlight the differences between LLM for similar mathematical topics. Another finding is that vague-language problems in mathematics may be different from problems with difficulty in solving the equations involved in the problem, as revealed by Calais et al. (2025). These results led to the conclusion that achievement in mathematics was dependent on a number of related skills and not just a general mathematical ability.

Another problem was made about the tie between the accuracy of the answers and validity of intermediate reason. The answer may include a legitimate change which it was not necessary to make; may be missing explanations or a calculation was incorrect but is not asked for in the conclusion. The correct and logical answer may still be given wrong if there is a slight math error. This work of Li et al. (2024) looked at the LLM ability to detect mathematical mistakes and resolve them, while stressing the need for evaluating reasoning outside of the answer itself. Further, Singh et al (2024) investigated mathematical error recognition and correction ability of language models. They were instrumental in advocating the recognition of conceptual, computational, representational and logical errors when evaluating. To the errors that could be logged, this information was also of great help with this type of

representation graph if only one solution was possible for the order of the transitions. This approach can help teachers or scientists determine which parts of grammar they and their students are struggling with, or which points are problematic for the A.I. system.

The study also presented suggestions for mathematics education and creating AI-enabled learning tools. As a result detailed solution trajectories might help teachers to learn how the students got from verbal to symbolic explanations, calculations and diagrams. They may also indicate if the students discovered procedures or made relations between mathematical ideas. For AI systems, representation graphs could be useful in clarifying the nature of the evaluation, and may help to reveal the steps that were accepted and which require further consideration. Zhai, Zhou and Ai (2024) demonstrated that giving feedback during intermediate points of mathematical reasoning as well as at the end of the mathematical reasoning, and/or the end of the composition is helpful to process supervision. The other side of that equation, Lightman et al. (2023) showed, is the process-based model of assessing mathematical solutions, which might be of assistance as well. These studies were encouraging about the potential of this type of representation graph, but failed to show the "mind" of the "thinking" participant humans and LLMs. The value of their use would therefore depend upon the careful recording and checking with mathematics, and the proper interpretation of the solutions so recorded.

The novelty of the study is that it integrated the representational analysis approach, and compared the human solution trajectory with that used by LLMs in two mathematical domains. The framework it proposed was able to link the accuracy of the answers, the diversity of representations, the validity of the transition, the level of complexity of the solution, and pattern of the errors in one analytical structure. There were a few things, however, that needed to be taken into account, such as a few limitations. The learners, models and mathematical tasks selected in the sample were 60 university students, 3 LLMs, and 20 problems, respectively, which is a small number of all learners, models and problems, respectively. Model answers can also vary in response to the wording in the prompting question, model versions and settings. The researchers may wish to test their model in the future in familiar environments, use different levels of education and a greater number of students. It can also investigate if representation graphs helped teachers to get a better understanding of students' misconceptions and whether these could be used to aid in checking explanations produced by AI. It helped to formulate a way to work mathematically towards solutions in an organized way, and gave the need to test further to see its educational value and reliability.

Conclusion

The study explored the study of mathematical reasoning by comparing the human and large language model (LLM) solution trajectories in algebra and geometry with the representation graph. It highlighted the problem solving skills with respect to using equations, numerical calculations and drawings along with verbal reasoning to solve the problem. The proposed structure allowed for a systematic investigation of the correctness of some answers, the variability of their representation, and their mathematical errors or validity of the transition to the problem's solution. The authors used the study to uncover types of problem-solving approaches, which aren't apparent through the end solutions. The results of reasoning were also highlighted as different to correct results; this was achieved in the comparison. The findings framework might facilitate mathematics educators to recognize learning challenges, and researchers more systematically review the mathematics explanation created by AI tools. More empirical research with trusted information, a variety of math problems, and various LLMs is needed to validate the trustworthiness and generalizability of representation graphs. The study not only introduced a process perspective into mathematical reasoning, but also provided a foundation for further research aimed at gaining insights into human cognition, mathematical representations, and the application of AI in mathematical education.

References

- Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 225–237. <https://doi.org/10.18653/v1/2024.eacl-srw.17>
- Ayyıldız Altınbaş, A., Solak, S., & Ertekin, E. (2025). The effect of multiple representations-based and problem-based instruction in linear algebra on pre-service teachers' dimensions of understanding and self-efficacy perceptions. Journal of Mathematics Teacher Education. <https://doi.org/10.1007/s10857-025-09696-0>
- Calais, P., Franco, G., Tang, Z., Nikas, T., Meira, W., Jr., Terzi, E., & Crovella, M. (2025). Disentangling text and math in word problems: Evidence for the bidimensional structure of large language models' reasoning. Findings of the Association for Computational Linguistics: ACL 2025. <https://aclanthology.org/2025.findings-acl.656/>

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv. <https://arxiv.org/abs/2110.14168>

Davoodi, A. G., Davoudi, S. P. M., & Pezeshkpour, P. (2025). LLMs are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of LLMs. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. <https://doi.org/10.18653/v1/2025.naacl-long.161>

Desai, S., Bush, S. B., & Safi, F. (2021). Mathematical representations in the teaching and learning of geometry: A review of the literature from the United States. Electronic Journal for Research in Science & Mathematics Education, 25(4), 6–22.

Gholami Davoodi, A., Mousavi Davoudi, S. P., & Pezeshkpour, P. (2025). LLMs are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of LLMs. Association for Computational Linguistics.

Jitendra, A. K., Dougherty, B., Sanchez, V., & Suchilt, L. (2024). Using multiple representations to foster multiplicative reasoning in students with mathematics learning disabilities. Teaching Exceptional Children, 57(2). <https://doi.org/10.1177/00400599221115627>

Li, C., Wang, W., Hu, J., Wei, Y., Han, H., Zhang, Z., & Peng, H. (2024). Common 7B language models already possess strong math capabilities. arXiv. <https://arxiv.org/abs/2403.04706>

Li, X., Wang, W., Li, M., Guo, J., Zhang, Y., & Feng, F. (2024). Evaluating mathematical reasoning of large language models: A focus on error identification and correction. Findings of the Association for Computational Linguistics: ACL 2024, 11316–11360. <https://doi.org/10.18653/v1/2024.findings-acl.673>

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's verify step by step. arXiv. <https://arxiv.org/abs/2305.20050>

-
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2024). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. International Conference on Learning Representations. <https://arxiv.org/abs/2310.02255>
- Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv. <https://arxiv.org/abs/2410.05229>
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., & Gao, M. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv. <https://arxiv.org/abs/2402.03300>
- Singh, J., Nambi, A., & Vineet, V. (2024). Exposing the Achilles' heel: Evaluating LLMs ability to handle mistakes in mathematical reasoning. Microsoft Research.
- Wang, P.-Y., Liu, T.-S., Wang, C., Wang, Y.-D., Yan, S., Jia, C.-X., Liu, X.-H., Chen, X.-W., Xu, J.-C., Li, Z., & Yu, Y. (2025). A survey on large language models for mathematical reasoning. arXiv. <https://arxiv.org/abs/2506.08446>
- Yan, Y., Su, J., He, J., Fu, F., Zheng, X., Lyu, Y., Wang, K., Wang, S., Wen, Q., & Hu, X. (2025). A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. Findings of the Association for Computational Linguistics: ACL 2025, 11798–11827. <https://doi.org/10.18653/v1/2025.findings-acl.614>
- Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Gao, P., & Li, H. (2024). MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? arXiv. <https://arxiv.org/abs/2403.14624>
- Zhang, Y., Xue, M., Liu, D., & He, Z. (2024). Rationales for answers to simple math word problems confuse large language models. Findings of the Association for Computational Linguistics: ACL 2024, 8853–8869. <https://doi.org/10.18653/v1/2024.findings-acl.524>
-

