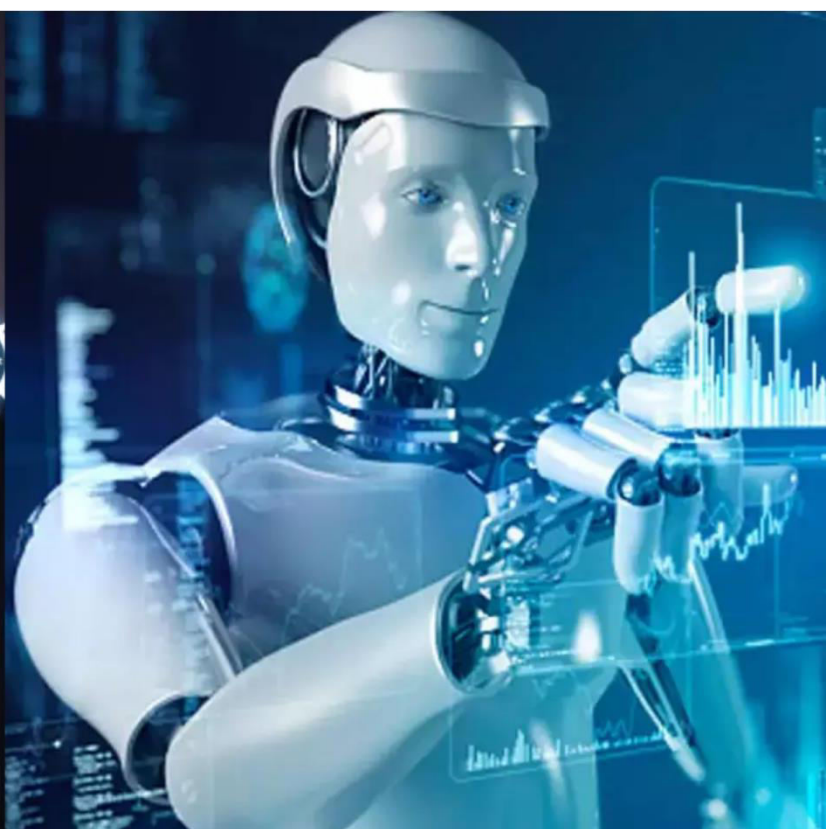




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 4, April 2026

LLMs for Mathematical Problem Solving: Tool Use, Step Validation, and Formal Checking as a Reliability Stack

Aarav Aggarwal

Student, Pathways School, Noida, Uttar Pradesh, India

ABSTRACT: Large Language Models (LLMs) have shown remarkable performance in solving mathematical problems, which makes them useful for educational purposes, research, and decision-making. Their abilities are frequently hampered by arithmetic errors, logical inconsistencies, hallucinated reasoning steps and errors in multi-step problem solving, however. This paper summarizes the contributions of tool use, step validation, and formal checking to a reliability stack for improving mathematical reasoning accuracy in LLMs. The secondary research methodology was adopted, where research journals, conference papers, and scholarly publications were analysed for their content. The review emphasizes the need for structured decomposition methods, external computational tools, and verification processes to minimize reasoning mistakes and enhance solution quality. The results show that none of these techniques can guarantee reliability, and only a combination of decomposition, tool augmentation, and verification will give the best framework. The research results show that a multi-layered reliability stack can greatly enhance the transparency, trustworthiness, and overall performance of LLMs for mathematical problem solving. The focus of the presentation is on Large Language Models (LLMs), Mathematical Reasoning and Verification Mechanisms.

KEYWORDS: Large Language Models (LLMs), Mathematical Reasoning, Verification Mechanisms

I. INTRODUCTION

The rise of Large Language Models (LLMs) has led to their notable achievements as strong artificial intelligence systems that can execute a wide variety of reasoning tasks, such as mathematical problem solving. There has been a recent progress of models like GPT, Gemini, Claude, and LLaMA in their ability to solve arithmetic, algebraic, and complex math problems (Zhou and Zhang, 2025). While these successes are impressive, LLMs are prone to mistakes in arithmetic, multiple-step reasoning, and in providing a logical step that is not actually in the text, as well as to fail to provide an answer to the question. However, whilst these successes are remarkable, LLMs also make errors, such as those in arithmetic, multiple step reasoning, and by inserting a step that is not logically present in the text, and by omitting an answer to the question. These restrictions have implications for their application in other areas of education, scientific and professional work, where mathematical correctness is important (Zheng et al., 2025). Researchers have come up with a layered design that integrates structured decomposition, external tool usage and verification mechanisms, to improve the reliability of mathematical reasoning. Structured decomposition techniques, such as Chain-of-Thought prompting, enable models to break complex problems into smaller and more manageable reasoning steps. Calculators, Computer Algebra Systems (CAS) and retrieval systems are used externally as accurate computing tools and for accessing relevant mathematical knowledge. In addition, verification techniques like self-consistency, step validation, verifier-guided generation and formal proof checking assists in identifying and rectifying errors in reasoning. In this paper, these strategies for increasing reliability are reviewed, common error modes in LLM mathematical reasoning are discussed, and various methods for mitigating the negative effects of these error modes on the accuracy, transparency, and trustworthiness of mathematical problem solving are compared (Qian et al., 2025).

II. RATIONALE AND OBJECTIVE

Rationale

Human-like reasoning and solving complex mathematical tasks have made the use of Large Language Models (LLMs) in mathematical problem solving an interesting focus. Despite their remarkable performance, however, LLM often produce wrong answers due to arithmetic error, logical error, hallucinated reasoning steps and inconsistencies in reasoning over multiple steps problem solving. These restrictions are limiting their products, especially in educational,

scientific and professional environments where precision is essential (Cao et al., 2025). As such, several approaches to enhance reliability have been proposed, such as structured decomposition techniques, external tools integration, and verification mechanisms. To create more accurate and trustworthy AI systems, it is crucial to grasp their effectiveness. This study intends, therefore, to investigate the enhancement of mathematical reasoning capabilities of LLM models as well as the reduction of the common error patterns (Chojecski, 2025).

Objectives

1. To examine how structured decomposition techniques, such as Chain-of-Thought and Tree-of-Thought reasoning, improve mathematical problem-solving performance in Large Language Models.
2. To investigate the role of external tools, including calculators, Computer Algebra Systems (CAS), and retrieval systems, in enhancing computational accuracy and reasoning capabilities.
3. To evaluate the effectiveness of verification mechanisms and mitigation strategies in reducing errors and improving the reliability of mathematical reasoning generated by Large Language Models.

III. METHODOLOGY

This research uses a qualitative secondary research approach to explore the role of tool use, step validation, and formal checking in improving the reliability of Large Language Models (LLMs) in mathematical problem solving. Secondary data were gathered from an extensive literature review of published academic research, such as peer reviewed journal articles, conference papers, scholarly books and research papers. Academic databases and search engines like Google Scholar were used to locate relevant publications that focus on mathematical reasoning, Chain-of-Thought prompting, tool-augmented LLMs, verification mechanisms, and formal proof systems (Guta and Kuser, 2025). The literature was searched for the key themes, common error modes and reliability enhancing approaches. The effectiveness of each of the following approaches: structured decomposition, integration of external tools, self-consistency, generation based on verifiers and formal checking frameworks is evaluated by comparative analysis. The results of various sources were combined to gain a general picture of the state of the art and research activities in the area. Ethical aspects were prudently handled throughout the study. No human subjects were involved, nor was personal data collected, in the production of all information, which was gathered from publicly available academic sources. Adequate referencing and referencing skills used to credit the original author of text and ideas to avoid plagiarizing. Research conducted in an honest and responsible way in adherence to academic honesty, transparency and responsible use of scholarly information (Petrov et al., 2026).

IV. LITERATURE REVIEW

Theme 1: Mathematical Reasoning Capabilities of Large Language Models

The field of Artificial Intelligence has undergone a major shift in the last few years with the rise of Large Language Models or LLM's that perform very well on a range of reasoning and language tasks. Recent models like GPT, Gemini, Claude and LLaMA have demonstrated significant progress in tackling mathematical tasks from simple computations to complex algebra and geometry problems (Liu et al., 2025). These models are fed with vast amounts of data that include textbooks, research papers, web pages, and mathematical examples, and then learn to recognize patterns and produce responses that are similar to human reasoning. There is evidence that larger sizes of the models and larger amounts of training data have helped improve mathematical reasoning performance. The larger models typically have been found to work more successfully on the benchmark data sets because they have the capacity to learn correlation between mathematical concepts and procedures of problem solving. Despite the progress made, math reasoning is still an elusive endeavor (Anand et al., 2026). Mathematics is different from natural language generation, because it involves calculations with precise results, logical consistency and step-by-step reasoning. There are multiple reports about the tendency of LLMs to produce plausible but wrong answers. The final answer often looks like it's correct, but there are often mistakes in the supporting arguments. Some of these errors involve errors in calculations, errors in premises or errors in reasoning. LLMs are believed to fall short in complex reasoning because they are mostly statistical pattern recognisers and lack mathematical understanding (Marinosyan, 2024). What's more, as a problem becomes more complex, its performance decreases. For multi-step mathematical problems, the model must keep in mind the relationship between the steps of the reasoning and its context. In fact, small mistakes made in the initial stages of the solution process can work their way through the solution and result in wrong conclusions. In light of this, researchers have turned to enhancing the reliability of reasoning, as opposed to making models larger.

This has resulted in the development of mathematical reasoning-based decomposition techniques, tools and verification mechanisms specifically tailored for mathematical reasoning problems (Huang et al., 2025).

Theme 2: Structured Decomposition Techniques for Mathematical Problem Solving

Structured Decomposition is one of the most extensively studied methods for enhancing mathematical reasoning in LLMs. This method prompts models to break down big problems into small and manageable ones and then produce a final answer. Decomposition is seen by researchers to be similar to human problem solving and to reduce cognitive complexity of reasoning. One of the most impactful decomposition methods is called chain of thought (CoT) (Chen, 2025). The model is asked to explain the way it makes its decision, rather than giving an answer right away. The study found that even when the model is instructed to execute intermediate steps in the reasoning process, CoT prompting has a significant impact on improving mathematical benchmarks, showing that the model can organize its intermediate reasoning before reaching a conclusion. This approach also allows the user to trace back through the reasoning to see where the mistakes have been made. Tree-of-Thought (ToT) prompting is a method that builds on Chain-of-Thought prompting. ToT is different from CoT, in the sense that it considers more than one possible solution path at the same time. This helps the model to consider other possibilities and make the best choice for reaching a correct answer (Cai et al., 2025). The results of the research show that using Tree-of-Thought reasoning results in better performance on some challenging math problems because it mitigates reliance on a single reasoning path that may be incorrect. Least-to-Most prompting, that is, solving a complex problem in steps of simpler ones, is another important technique. The solution of every sub-problem is done separately and then added to the total solution (Metere, 2026). This approach has been especially successful for tasks that involve multiple steps of reasoning due to the ease of not overloading the model with too many steps. Although the decomposition methods have their benefits, they are not without their drawbacks (Gómez-Martín and Álvarez, 2026). Even when an intermediate step is incorrect, subsequent reasoning can still be impacted. Also, the longer the reasoning chains, the more compute is required and processing time. However, structured decomposition is a key part of developing reliable mathematical reasoning, and is the lowest level in the reliability stack (Carbone, 2025).

Theme 3: Tool-Augmented Reasoning and External Computational Support

The ability of LLMs to perform well on mathematical tasks alone is starting to be acknowledged as insufficient for precision. Given this, extensive efforts have been dedicated to incorporating external computation capabilities that augment language model reasoning capabilities. One of the easiest, and most effective, methods is integration with a calculator (Ye et al., 2025). Many mathematical problems may require arithmetic calculations that can be better done with a maths-specific computing system than with a neural network. Delegating calculations to calculators can help LLM minimize numerical errors and engage in more abstract thought problems. Research shows that models that use a calculator can outperform language models on math benchmarks. The more sophisticated level of tool support is given by the so-called Computer Algebra Systems (CAS). These systems can conduct symbolic mathematical calculations such as differentiation, integration, solving of equations and matrix calculations. Whereas standard calculators compute values of expressions, CAS platforms act symbolically on the expressions, which makes them very useful in solving complex mathematical problems. Research indicates that embedding CAS capabilities into LLMs significantly enhances their capabilities on complex mathematical reasoning tasks. Another key advancement is the use of Retrieve-Augmented Generation (RAG) (Josephson, 2025). This method allows models to retrieve external knowledge bases when solving a problem. Rather than just memorizing information, LLMs can pull up formulas, definitions, theorems and relevant examples from other sources. They have discovered that retrieval systems decrease hallucinations and enhance factual accuracy, especially in solving domain specific mathematical problems. There are several benefits of using tool augmentation. It increases the accuracy of computation, helps to avoid arithmetic errors, and makes it easier to access specialized knowledge. But issues are still left to be addressed in tool selection, coordination and integration (Wang et al., 2025). A wrong reasoning process will not necessarily lead to a wrong use of the tools and hence to wrong conclusions. So, tool support is not sufficient for reliability and needs to be complemented by verification mechanisms to achieve maximum effectiveness.

Theme 4: Verification Mechanisms and Step Validation

Verification mechanisms are an important part of the reliability stack as they are concerned with checking the correctness of solutions produced. Researchers say verification is needed because of the potential for error even in sophisticated reasoning and with the help of tools. Self-consistency is one of the most popular verification techniques (Wang et al., 2026). This method creates several different lines of reasoning for the same problem and picks out the

one most common to all solutions. The premise is that there are more probable correct answers that occur repeatedly, and there are less probable incorrect answers that occur repeatedly. Empirical research shows self-consistency to have a measurable effect on the accuracy of mathematical reasoning; self-consistency reduces the effect of random reasoning errors. Another approach is step validation. Step validation considers each step of the reasoning process individually rather than just the end result. Individual transitions are evaluated in a dedicated validation model or with a rule-based system, to determine logical consistency (Guo et al., 2025). This way, errors can be identified as early as they can be in the reasoning process, before they are carried over to the next steps in the calculation. One of the potential research areas is verifier guided generation. Within this context, a verifier model can be used to assess candidate solutions generated by an LLM. The verifier checks the logical consistency, mathematical correctness and fulfilment of the problem requirement. Solutions which have been verified with higher scores are given priority or chosen as final products. Studies have shown that verifier-guided methods can significantly enhance the reliability of results, especially in more complex mathematical problems (Saha et al., 2025). Formal proof verification is the most stringent method of proof validation. Formal systems like Lean, Coq, Isabelle test mathematical statements using rigorous logic rules. A proof must be formally justified and for this, every reasoning step must be formally justified. Formal verification is a very high confidence approach to correctness but still uses significant resources and expertise. In summary, verification mechanisms can be viewed as crucial tools in mathematical reasoning systems. They play an important role in enhancing trust in AI-generated solutions and detecting hidden errors, and aid in boosting confidence in solutions (Wang et al., 2026).

Theme 5: Reliability Challenges and Future Directions in Mathematical Reasoning

While there has been great progress, several challenges still exist for LLM reliability in mathematical problem solving (Wen et al., 2024). These four types of errors—arithmetic, logical, hallucination, and context management—have consistently been identified as significant challenges for building trustworthy AI systems (Buali, 2025). Even at the highest levels of language models, arithmetic errors still happen. These models are good at capturing accurate meaning of mathematical concepts, but may lack the ability to make precise calculations. Errors of logical reasoning are also made if models make assumption(s) that are not supported or if they are using mathematical principles incorrectly. These types of errors have the potential to undermine entire solution pathways, and lower overall reliability. Another constant difficulty is hallucination. In some cases, LLM-generated intermediate steps seem reasonable but are not mathematically correct (Zhang et al., 2026). Users may be misled into believing that they are steps, and may reach the wrong conclusions. Reasoning tasks are further complicated by context management problems, such as when problems involve holding information across a number of steps in analysis. More and more researchers are suggesting hybrid systems, which incorporate decomposition methods, external tools, and verification methods. These systems mimic human mathematical practice: divide large problems into smaller parts, use reliable tools to do the calculations, and check the results before they are accepted. There is some evidence that an integrated approach is superior to the individual approaches because of the ability to reduce more than one source of error at a time. In the future, further work is anticipated to be done on the tighter integration between language models and formal reasoning systems (Hao et al., 2025). New developments in automated theorem prover, symbolic reasoning, and verifier-guided learning can make mathematical reliability even better. Also, transparency and interpretability will continue to be important to build trust with users. With the growing applications of mathematical reasoning across the education, research, and decision-making domains, the construction of comprehensive reliability schemes will remain a key priority for the AI community (Azerbaiyev et al., 2023).

Summary

In general, the current literature contains two stories about AI-based fintech platforms. On the one hand, they contribute to the improvement of financial decision-making by automation, personalisation, and the minimisation of behavioural bias. Conversely, they come with new challenges associated with behavioural nudging, algorithmic dependence, insufficient transparency, and less autonomy of investors. It is also evident in the literature that AI in fintech is not a passive facilitator of financial behaviour but a contributor to it. This generates a significant gap in the research on whether or not algorithmic systems have a subtle and real power to empower or manipulate retail investors.

V. CONCLUSION

The studies reviewed in this work underscore the increasing potential of Large Language Models (LLMs) in mathematics problem-solving and the reliability and accuracy issues that arise with them. Though sophisticated LLMs excel at arithmetic, algebraic, and reasoning-based mathematical problems, they can still be prone to errors like incorrect answers, erroneous reasoning steps, and missing steps in multi-step tasks. While powerful LLMs can handle arithmetic, algebraic, and reasoning problems with effectiveness, they are subject to problems such as incorrect answers, missing steps in multi-step problems, and errors in reasoning steps. Such restrictions suggest that just by using a model and training it, a reliable mathematical reasoning cannot be ensured. The review identified 5 key themes. First, LLMs have demonstrated excellent mathematical reasoning abilities and remain prone to reliability problems. Second, structured decomposition methods like Chain-of-Thought (COT) and Tree-of-Thought (ToT) enhance reasoning by dividing problems into smaller parts. Thirdly, external tools such as retrieval devices and Computer Algebra Systems, and calculators improve the accuracy of computations and access to knowledge. Fourth, verification techniques including self-consistency, step validation, verifier guided generation and formal checking can be used to detect and minimize reasoning errors. Lastly, current reliability issues point to the need for integrated solutions, involving reasoning, tools and verification. In conclusion, the literature indicates that the most effective approach to augment mathematical problem solving in LLMs is a reliability stack comprising of decomposition, tool augmentation, and verification. Continued work will be aimed at developing better, more transparent, more reliable mathematical reasoning AI systems.

REFERENCES

1. Anand, A., Ramesh, M., Mittal, A., Kumar, A., Cambria, E., Wang, Z., Liu, T., Ng, A.B., See, S. and Shah, R.R., 2026. The Periodic Table of LLM Reasoning: A Structured Survey of Reasoning Paradigms, Methods, and Failure Modes. arXiv preprint arXiv:2606.11470.
2. Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M.D., McAleer, S., Jiang, A.Q., Deng, J., Biderman, S. and Welleck, S., 2023. Llemma: An open language model for mathematics. arXiv preprint arXiv:2310.10631.
3. Buali, M., 2025. Automated Theorem Proving with Large Language Models in Lean: An Exploration of Specialized In-Context Learning and General-Purpose Hierarchical Architectures (Doctoral dissertation, King Abdullah University of Science and Technology).
4. Cai, Y., Hou, Z., Sanán, D., Luan, X., Lin, Y., Sun, J. and Dong, J.S., 2025. Automated program refinement: Guide and verify code large language model with refinement calculus. Proceedings of the ACM on Programming Languages, 9(POPL), pp.2057-2089.
5. Cao, J., Lu, Y., Li, M., Ma, H., Li, H., He, M., Wen, C., Sun, L., Zhang, H., Qin, S. and Cheung, S.C., 2025, July. From informal to formal-incorporating and evaluating llms on natural language requirements to verifiable formal proofs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 26984-27003).
6. Carbone, L., 2025. Advancing mathematics research with generative AI. arXiv preprint arXiv:2511.07420.
7. Chen, S., 2025. Toward Reliable Problem Solving with Pre-trained Large Language Models (Doctoral dissertation, University of Toronto (Canada)).
8. Chojacki, P., 2025. Mathematics in the Age of Large Language Models. preprint at ulam. ai/research.
9. Gómez-Martín, F. and Álvarez, A.A., 2026. Can Large Language Models Generate Higher-Order Thinking Math Problems? An Evaluation Study. IEEE Access.
10. Guo, J., Huang, S., Li, M., Huang, D., Chen, X., Zhang, R., Guo, Z., Yu, H., Yiu, S.M., Lio, P. and Lam, K.Y., 2025. A Comprehensive Survey on Benchmarks and Solutions in Software Engineering of LLM-Empowered Agentic System. arXiv preprint arXiv:2510.09721.
11. Guta, G. and Kuser, G., 2025. LLM-based framework to support the construction of valid formal models.
12. Hao, Y., Chen, Y., Zhang, Y. and Fan, C., 2025, April. Large language models can solve real-world planning rigorously with formal verification tools. 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) (pp. 3434-3483).
13. Huang, L., Mustafin, I., Piccioni, M., Schena, A., Weber, R. and Meyer, B., 2025. Do AI models help produce verified bug fixes?. arXiv preprint arXiv:2507.15822.

14. Josephson, C., 2025. Proof-Carrying Verification for Tool-Using LLM Agents: Frozen Records, One-Clock Certificates, and Refinement Ladders. One-Clock Certificates, and Refinement Ladders (December 31, 2025).
15. Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., He, M., Chen, Q., Jiang, B., Zhou, A. and He, L., 2025. Mathematical language models: A survey. ACM Computing Surveys, 58(6), pp.1-37.
16. Marinosyan, A., 2024. Llms in physics and mathematics education and problem solving: Assessment of chatgpt-4 level and suggestions for improvement.
17. Metere, A., 2026. Methods for Formal Verification of Agent Skills: Three Layers Toward a Mechanically Checkable Capability-Containment Proof. arXiv preprint arXiv:2605.23951.
18. Petrov, I., Dekoninck, J., Dimitrov, D.I. and Vechev, M., 2026. Not All Proofs Are Equal: Evaluating LLM Proof Quality Beyond Correctness. arXiv preprint arXiv:2605.10379.
19. Qian, C., Du, H., Wang, H., Chen, X., Zhang, Y., Sil, A., Zhai, C., McKeown, K. and Ji, H., 2025, May. ModelingAgent: Bridging LLMs and Mathematical Modeling for Real-World Challenges. In Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 1599-1633).
20. Saha, D., Tarek, S., Shaikh, H.A., Hasan, K.T., Nalluri, P.S., Hasan, M.A., Alam, N., Zhou, J., Saha, S.K., Tehranipoor, M. and Farahmandi, F., 2025. Sv-llm: An agentic approach for soc security verification using large language models. arXiv preprint arXiv:2506.20415.
21. Wang, E.Y., Motwani, S., Roggeveen, J.V., Hodges, E., Jayalath, D., London, C., Ramakrishnan, K., Cipcigan, F., Torr, P. and Abate, A., 2026. Horizonmath: Measuring AI progress toward mathematical discovery with automatic verification. arXiv preprint arXiv:2603.15617.
22. Wang, J., Ming, Y., Ke, Z., Xiong, C., Joty, S., Albarghouthi, A. and Sala, F., 2026. Beyond accuracy: Dissecting mathematical reasoning for llms under reinforcement learning. Advances in Neural Information Processing Systems, 38, pp.66616-66653.
23. Wang, S., Hu, X., Chen, J., Pan, Z. and Xia, X., 2025. Open the Oyster: Empirical Evaluation and Improvement of Code Reasoning Confidence in LLMs. arXiv preprint arXiv:2511.02197.
24. Wen, C., Cao, J., Su, J., Xu, Z., Qin, S., He, M., Li, H., Cheung, S.C. and Tian, C., 2024, July. Enchanting program specification synthesis by large language models using static analysis and program verification. In International Conference on Computer Aided Verification (pp. 302-328). Cham: Springer Nature Switzerland.
25. Ye, J., Liu, T., Tian, Q., Su, S., Jiang, Z. and Wang, X., 2025. Chatmodel: Automating reference model design and verification with llms. arXiv preprint arXiv:2506.15066.
26. Zhang, F., Xu, J., Wang, C., Cui, C., Liu, Y. and An, B., 2026. Incentivizing llms to self-verify their answers. Advances in Neural Information Processing Systems, 38, pp.4800-4829.
27. Zheng, D., Du, L., Su, J., Tian, Y., Zhu, Y., Zhang, J., Wei, L., Zhang, N. and Chen, H., 2025. Knowledge augmented complex problem solving with large language models: A survey. arXiv preprint arXiv:2505.03418.
28. Zhou, K. and Zhang, L., 2025. Step-wise formal verification for llm-based mathematical problem solving. arXiv preprint arXiv:2505.20869.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details