Analisis Performa Teknik Prompting pada ChatGPT dalam Menyelesaikan Soal Kalkulus Menggunakan *Logistic Mixed Effects Model*

Tugas Akhir

diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana

dari Program Studi S1 Informatika

Fakultas Informatika

Universitas Telkom

1301213273

Cakra Budiman Putra

Program Studi Sarjana S1 Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2025

LEMBAR PENGESAHAN

Analisis Performa Teknik Prompting pada ChatGPT dalam Menyelesaikan Soal Kalkulus Menggunakan *Logistic Mixed Effects Model*

Performance Analysis of Prompting Technique on ChatGPT in Solving Calculus Problems Using Logistic Mixed Effects Model

1301213273

Cakra Budiman Putra

Tugas akhir ini telah diterima dan disahkan untuk memenuhu sebagai syarat memperoleh gelar pada Program Studi Sarjana S1 Informatika

Fakultas Informatika

Universitas Telkom

Bandung, 13 Agustus 2025

Menyetujui

Pembimbing I,	Pembimbing II,
Dr. Kemas Muslim L	**Dr. Aditya Firman Ihsan**
NIP: 13820075	NIP: 21950003

Ketua Program Studi
Sarjana S1 Informatika,

Mahmud Dwi Sulistiyo, S.T., M.T.,Ph.D.
NIP: 13880017

LEMBAR ORISINALITAS

Dengan ini saya, Cakra Budiman Putra, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul "Analisis Performa Teknik Prompting pada ChatGPT dalam Menyelesaikan Soal Kalkulus Menggunakan *Logistic Mixed Effects Model"* berserta dengan seluruh isinya merupakan hasil karya saya sendiri, dengan tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang berlaku dengan masyarakat keilmuan, serta produk dari tugas akhir ini bukan merupakan hasil dari *Generative AI.* Saya siap menggung risiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam Laporan Tugas Akhir, atau jika ada klai mdari pihak lain terhadap keaslian karya.

Bandung, 13 Agustus 2025

Yang menyatakan

Cakra

Cakra Budiman Putra

NIM: 1301213273

Performance Analysis of Prompting Technique on ChatGPT in Solving Calculus Problems Using Logistic Mixed Effects Model

1st Cakra Budiman Putra
School of Computing
Telkom University
Bandung, Indonesia
cakrabudiman@student.telkomuniversity.ac.id

2nd Kemas Muslim L
School of Computing
Telkom University
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

3rd Aditya Firman Ihsan
School of Computing
Telkom University
Bandung, Indonesia
adityaihsan@telkomuniversity.ac.id

Abstract— **The utilization of Large Language Models (LLMs) like ChatGPT in calculus education offers significant potential, yet its effectiveness is highly dependent on the prompting technique employed. This study aims to systematically evaluate and compare the performance of three prompting techniques Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) in solving calculus problems (integrals, derivatives, and limits) using the GPT-4o mini model. A total of 270 responses from 90 problems of varying difficulty levels were manually evaluated by experts based on four criteria: clarity, correctness, strategy, and representation. Data analysis was performed using a Logistic Mixed Effects Model (LMM) to test the influence of each variable. The results consistently demonstrate that the Chain-of-Thought (CoT) technique is significantly superior, particularly in enhancing the clarity and strategic quality of the solutions. The analysis also reveals that problem type is a critical factor, with limit problems posing the greatest challenge to the model. This study concludes that the choice of prompting technique is crucial for maximizing the potential of LLMs in education and recommends CoT as the most effective strategy for generating solutions that are not only accurate but also possess high pedagogical value.**

Keywords—ChatGPT, Prompting Techniques, Calculus Education, Artificial Intelligence, LLM, Logistic Mixed Effects Model (LMM)

I. Introduction

A. Background

The development of artificial intelligence (AI) technology has become a new opportunity in various fields, including education. One prominent AI application is ChatGPT, a Large Language Model (LLM). ChatGPT is able to process and generate text in various contexts. ChatGPT has been widely used to support learning, from helping students with assignments to automatically generating questions for educational assessments [1][2].

Especially in mathematics, calculus is known as a complex branch of science, including integrals, derivatives, and limits, which are often challenging for students. AI models like ChatGPT have great potential to help solve this problem by generating structured and logical solutions. [3][4].

Although many studies have shown that prompting techniques such as Chain-of-Thought (CoT) have proven to enhance ChatGPT's performance in solving difficult mathematical problems [5], this technique often requires longer and more complex prompt designs, which can be a challenge for general users [1][6]. This creates a gap between the potential of the technology and its ease of use in real-world learning scenarios. Furthermore, the use of free AI models like GPT-4o mini benefits everyone, as it can be accessed by various groups without additional costs. However, there is still a gap in its effectiveness. Therefore, an in-depth analysis is needed to systematically compare the performance of various prompting techniques to ensure that the generated solutions are not only accurate but also clear and relevant [1][7].

To comprehensively evaluate the factors affecting solution quality, this study uses a Logistic Mixed Effects Model (LMM). This method was chosen for its superior ability to analyze the complex relationships between various variables such as prompting techniques, problem types, and difficulty levels simultaneously. By considering the influence of fixed effects (consistent effects) and random effects (unique variations among problems), LMM can provide a more in-depth analysis to identify the determining factors of prompting effectiveness in AI-based calculus learning [8][9].

B. Topic and Its Limitations

This research focuses on evaluating the performance of three prompting techniques Zero-Shot, Few-Shot, and Chain-of-Thought in solving calculus problems using the ChatGPT model (GPT-4o mini version). The central issue under investigation is whether the choice of prompting technique affects the quality of the solution generated by the AI. The system's input is a calculus problem covering integrals, derivatives, or limits, which is structured according to a specific prompt format for each technique[10]. The output is the corresponding solution from ChatGPT, which is then manually assessed for its quality based on four criteria: clarity, correctness, strategy, and representation.

To maintain a clear focus within this defined scope, the study was designed with several key limitations. First, the problem material was confined to the three aforementioned calculus topics. Second, the dataset comprised 90 problems sourced from a single textbook, yielding 270 responses for analysis. Third, the evaluation was conducted manually by expert calculus lecturers at Telkom University to facilitate an in-depth qualitative assessment. Finally, to ensure variable consistency, this study exclusively used the GPT-4o mini model and did not perform a comparative analysis against other LLMs.

While these limitations were essential for methodological rigor, they also illuminate valuable directions for future research. These boundaries provide a clear foundation for studies aimed at exploring the contextual adaptability of prompting techniques more broadly. For instance, a future comparative analysis across different LLMs (e.g., GPT-4o mini vs. GPT-4o) could test the generalizability of the findings. Second, employing more varied problem types, such as word problems or applied scenarios, would assess the robustness of each technique in less structured contexts. A third avenue involves expanding the evaluation to include students as assessors to measure the pedagogical effectiveness of the solutions, not just their technical accuracy. Finally, applying this methodology to other mathematical domains, such as linear algebra or statistics, would help validate whether these results hold true for technical problem-solving in general.

C. *Objective*

Overall, this study aims to analyze and compare the performance of Zero-Shot, Few-Shot, and Chain-of-Thought prompting techniques on ChatGPT in the context of solving calculus problems. A further objective is to qualitatively evaluate the quality of the generated solutions based on the criteria of clarity, correctness, strategy, and representation, and to identify whether the type of calculus problem (integral, differential, and limit) affects the effectiveness of each techniqueFurthermore, this research applies the Logistic Mixed Effects Model (LMM) to explain the variability of solution quality influenced by the interaction between prompting techniques and problem types [11]. Ultimately, the main goal of this study is to provide evidence-based recommendations on the most effective and efficient prompting techniques for use in AI-supported calculus learning.

D. *Paper Organization*

This paper is organized as follows. After this Introduction section, a Related Studies section is presented, which reviews relevant previous research. The Research Methodology section describes in detail the experimental design, data collection, and evaluation process. Next, the Results and Analysis section presents the research findings, including the results of the statistical analysis using LMM. Last, the Conclusion section summarizes all findings and provides suggestions for next research.

II. Related Work

The utilization of Large Language Models (LLMs) like ChatGPT in mathematics education has become a focus of various studies. These studies consistently demonstrate that the effectiveness of LLMs is highly dependent on the interaction strategy used, also known as the prompting technique [12]. The most foundational research in this context is the introduction of Chain-of-Thought (CoT) Prompting by Wei et al. (2022), which showed that by encouraging the model to generate a series of intermediate reasoning steps before providing a final answer, its ability to solve complex mathematical problems can be significantly improved [5]. Wang et al. (2023) further confirmed that the effectiveness of CoT is heavily influenced by the design and context of the given prompt [13].

Different prompting techniques have distinct advantages and disadvantages depending on task complexity. The Zero-Shot technique, where the model is asked to complete a task without prior examples, and the Few-Shot technique, where a few examples are provided to guide the model, have proven adequate for simple problems in previous research [14]. However, for more intricate problems such as advanced calculus, these two techniques are often suboptimal [13]. Few-Shot prompting can serve as an effective bridge when the model needs to understand new patterns or contexts, although its success depends on the selection of representative examples [6]. On the other hand, CoT consistently excels in tasks requiring problem decomposition and step-by-step logical reasoning, making it a prime candidate for solving complex integral or derivative problems [5][7].

To measure the success of this technique, the evaluation process must be structured. Schorcht et al. (2024) highlighted four key criteria for assessing the quality of AI-generated mathematical solutions: clarity, correctness, strategy, and representation. Clarity refers to the readability and logical flow of the solution steps, where CoT has been proven effective in enhancing it [1]. Correctness is the mathematical accuracy of the final answer, while strategy evaluates the validity and efficiency of the approach taken. Lastly, representation, such as the use of graphs or symbols, can significantly support conceptual understanding, although some other LLMs do not support the use of mathematical symbols [3].

Given the often hierarchical structure of data in educational research (e.g., student responses nested within problems or classes), advanced statistical analysis methods are required[15]. Schorcht et al. (2024) emphasized the importance of using Logistic Mixed Effects Models (LMM) to analyze the relationships between variables like prompting techniques and solution quality [1]. LMMs can model binary data (e.g., correct/incorrect solutions) while accounting for fixed effects, such as prompt type or problem difficulty, and random effects, such as the inherent variability within each problem [8][9]. The ability of LMMs to handle complex data structures and analyze interactions between variables makes them a highly suitable tool for accurately evaluating the effectiveness of AI-based educational interventions [16][11].

III. Methodology

This study employs a structured experimental approach to evaluate and compare the performance of various prompting techniques on ChatGPT. The research methodology includes several key stages, from research design, data collection and evaluation, to statistical analysis using a Logistic Mixed Effects Model.

A. *Research Design and Flow*

The research flow involved collecting calculus problems, designing prompts, generating responses with ChatGPT, manual evaluation by experts, and finally, data analysis using an LMM to draw conclusions, as illustrated in Figure 1.

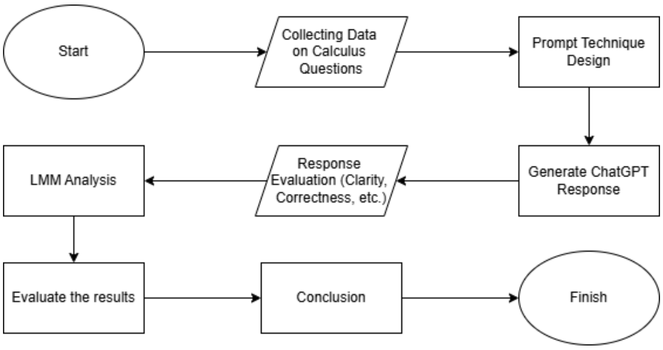
Start
Collecting Data on Calculus Questions
Prompt Technique Design
Generate ChatGPT Response
Response Evaluation (Clarity, Correctness, etc.)
LMM Analysis
Evaluate the results
Conclusion
Finish

Figure 1 Research System Flowchart

B. *Dataset dan Sumber Data*

The research dataset consists of 90 calculus problems sourced from the calculus textbook by Purcell [17]. These problems are classified into three main topics. Each topic includes problems with three difficulty levels. The structure of the problem dataset is presented in Table 1.

Table 1 Dataset Example

ID	Type	Description	Difficulty
1	Integral	Calculate the integral of ∫(2x^3 - x) dx	Easy
2	Derivative	Find the derivative of f(x) = e$^{(2x)}$ x sin(x)	Medium
3	Limit	Calculate the limit lim (x → ∞) (x^2 + 3)/(2x^2 - 5x + 6)	Hard

C. Experiment Procedure

The experiment was conducted by providing 90 calculus problems to the ChatGPT model, version GPT-4o mini. Each problem was tested using three different prompting techniques:

1) Zero-Shot Prompting: The problem was given directly without any examples.
2) Few-Shot Prompting: Several examples of problems and their solutions were provided before the main problem.
3) Chain-of-Thought (CoT) Prompting: Prompt with steps that are in accordance with your wishes or according to the rules[18].

This process will produce 270 responses (90 questions × 3 techniques) and will be used as data to be evaluated.

D. Evaluation System and Criteria

The evaluation of the responses was performed manually by evaluators, consisting of expert lecturers in the field of calculus at Telkom University. Manual assessment was chosen to ensure an in-depth analysis of qualitative aspects such as clarity and strategy. Each response was rated against four main criteria using a binary scale (1 for met, 0 for not met). The scoring criteria are summarized in Table 2.

Table 2 Scoring Criteria

Criteria	Description	Score
Clarity	Is the solution clear and easy to understand	0/1
Correctness	Is the answer correct or not	0/1
Strategy	Is the strategy is effective or not	0/1
Representation	There is a relevant and helpful visual or alternative representation	0/1

E. Research Variables

The variables used in this study are defined as follows:

1) Independent Variables
   a) Prompting Technique: Zero-Shot, Few-Shot, Chain-of-Thought).
   b) Problem Type: Integral, Derivative, Limit.
   c) Difficulty Level: Easy, Medium, Hard.
2) Dependent Variables : Binary score (0 or 1) for each evaluation criterion: Clarity, Correctness, Strategy, Representation.
3) Control Variables : Model Version ChatGPT GPT-4o mini.

F. Data Analysis

The binary data from the evaluation was analyzed using a Logistic Mixed Effects Model (LMM) to test the influence of the independent variables on each dependent variable. The LMM was chosen for its superior ability to model data with a hierarchical structure and to handle variability among problems [19]. To ensure the reliability and validity of the model's results, several validation steps were implemented. The statistical significance of each predictor coefficient was determined using a 95% Confidence Interval (CI), where an effect is considered significant if its range does not include zero. In addition to statistical validation, data reliability was also maintained during the manual evaluation phase; inter-rater reliability was ensured through discussion and calibration of the evaluation criteria among the expert assessors to minimize variability.

In this validated model, the prompting technique, problem type, and difficulty level are treated as fixed effects, while the unique identity of each problem is treated as a random effect. The general formula for an LMM can be written as follows [16]:

$$logit\big(P(y_{it} = 1)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + b_j + \epsilon_{it}$$

In the context of this research, this formula is adapted to analyze the influence of fixed effects (prompting technique, problem type, difficulty level) on solution quality, while controlling for the random effect arising from the inherent variation among problems[20]. Thus, the specific model used is:

$$logit\big(P(y_{it} = 1)\big) = \beta_0 + \beta_1(Prompting) + \beta_2(JProblem\ Type) + \beta_3(Difficulty) + (1|Problem\ ID)$$

Where:

1) (P(y$_{it}$=1)): The probability that a solution meets the evaluation criterion (clarity, correctness, strategy, representation) for the i-th problem assessed by the t-th evaluator.
2) logit(P): The log-odds of that probability. The logit function is used to transform the probability (0/1) into a continuous value.
3) β_0 : The intercept (constant).
4) β_1(Prompting): The fixed effect coefficient for the prompting technique (COT,FS,ZS)
5) β_2 (Problem Type): The fixed effect coefficient for the calculus problem type (integral, differential, limit).
6) β_3 (Difficulty): The fixed effect coefficient for the problem's difficulty level (easy, medium, hard).
7) (1 | Problem ID) : The random effect representing the variability between problems.

IV. Evaluation

This chapter presents the results of the experiments that have been conducted. The presentation begins with a descriptive analysis to provide a general overview of ChatGPT's performance based on three main variables: prompting technique, problem type, and difficulty level. Subsequently, the results of the inferential analysis using the Logistic Mixed Effects Model (LMM) are presented to test the significance of each variable's influence on the four solution quality criteria. This chapter concludes with a discussion interpreting these findings.

The analyzed dataset consists of 270 ChatGPT responses, derived from 90 calculus problems (integral, derivative, limit) with three difficulty levels (easy, medium, hard), each tested using three prompting techniques (Zero-Shot, Few-Shot, and Chain-of-Thought).

A. Descriptive Analysis

1) Mean Scores by Prompting Technique

Initial analysis shows that the prompting technique has a clear influence on solution quality. Overall, Chain-of-Thought (CoT) proved to be the most effective technique. CoT significantly excelled in producing solutions that were easy to understand (Clarity 0.93) and used the best solution steps (Strategy 0.93). Conversely, although the Few-Shot technique was able to produce accurate answers (Correctness 0.96), its explanations tended to be unclear (Clarity 0.71). Meanwhile, Zero-Shot showed reasonably accurate performance but was deemed to use less efficient strategies (Strategy 0.83).

Table 3 Prompting Technique Mean Score

Technique	Clarity	Correctness	Strategy	Representation
COT	0.93	0.97	0.93	1.00
Few-Shot	0.71	0.96	0.84	0.99
Zero-Shot	0.86	0.96	0.83	1.00

2) Mean Scores by Problem Type

ChatGPT's performance also varied depending on the type of calculus problem. The model demonstrated near-perfect mastery on Integral problems across all evaluation criteria. The greatest challenge arose with Limit problems, where the model often failed to select the appropriate solution strategy (Strategy 0.62), which consequently lowered the correctness of the answers. An interesting finding was observed for Derivative problems, where ChatGPT provided 100% correct answers with a perfect strategy but failed to deliver clear explanations (Clarity 0.59).

Table 4 Problem Type Mean Score

Technique	Clarity	Correctness	Strategy	Representation
Integral	0.99	0.99	0.99	0.99
Limit	0.92	0.89	0.62	1.00
Turunan	0.59	1.00	1.00	1.00

3) Mean Scores by Difficulty Level

While medium problems produced balanced results, hard problems revealed a clear trade-off. The model achieved excellent strategy and correctness (0.97) but with significantly lower clarity (0.77), indicating it prioritizes a correct answer over a clear explanation for complex tasks.

Table 5 Difficulty Level Mean Score

Technique	Clarity	Correctness	Strategy	Representation
Easy	0.82	0.96	0.76	1.00
Medium	0.91	0.96	0.89	0.99
Hard	0.77	0.97	0.97	1.00

B. Results of the Logistic Mixed Effects Model (LMM)

For the inferential analysis, four Logistic Mixed Effects Models (LMMs) were used, with Chain-of-Thought, Integral problems, and Easy difficulty set as the baseline categories. An effect was considered significant if its 95% CI excluded zero.

1) Clarity

Showed that both the Few-Shot technique and Derivative problems caused a significant decrease in solution clarity compared to their baselines.

Table 6 Coefficient for Clarity

Variable	Coef	CI Lower	CI Upper
Intercept	1.078	0.979	1.177
Technique[T.Few-Shot]	-0.222	-0.312	-0.132
Technique[T.Zero-Shot]	-0.078	-0.168	0.012
ProblemType[limit]	-0.067	-0.157	0.024
ProblemType[derivative]	-0.400	-0.490	-0.310
Difficulty[medium]	0.089	-0.001	0.179
Difficulty[hard]	-0.056	-0.146	0.035
Group Var	0.001		

2) Correctness

For answer correctness, the only factor with a significant influence was the problem type, where Limit problems showed a significant decrease in correctness compared to Integral problems. No significant differences were found between prompting techniques for this criterion.

Table 7 Coefficient for Correctness

Variable	Coef	CI Lower	CI Upper
Intercept	0.993	0.931	1.055
Technique[T.Few-Shot]	-0.011	-0.066	0.044
Technique[T.Zero-Shot]	-0.011	-0.066	0.044
ProblemType[limit]	-0.100	-0.155	-0.045
ProblemType[derivative]	0.011	-0.044	0.066
Difficulty[medium]	-0.000	-0.055	0.055
Difficulty[hard]	0.011	-0.044	0.066
Group Var	0.001		

3) Strategy

The model indicates that the Few-Shot and Zero-Shot techniques significantly lowered the strategy quality compared to CoT. Limit problems also significantly decreased the strategy score. Interestingly, medium and hard difficulty problems significantly increased strategy quality, suggesting that a greater challenge prompts more structured solutions.

Table 8 Coefficient for Strategy

Variable	Coef	CI Lower	CI Upper
Intercept	0.937	0.852	1.022
Technique[T.Few-Shot]	-0.089	-0.168	-0.010
Technique[T.Zero-Shot]	-0.100	-0.179	-0.021
ProblemType[limit]	-0.367	-0.445	-0.288
ProblemType[derivative]	0.011	-0.068	0.090
Difficulty[medium]	0.133	0.055	0.212
Difficulty[hard]	0.211	0.132	0.290
Group Var	0.000		

4) Representation

For the representation criteria, no predictor variables showed a significant effect. This indicates that the model's ability to provide visual or symbolic representations is relatively consistent and is not significantly affected by the instruction technique, problem type, or difficulty level in this study.

Table 9 Coefficient for Representation

Variable	Coef	CI Lower	CI Upper
Intercept	0.996	0.977	1.015
Technique[T.Few-Shot]	-0.011	-0.029	0.006
Technique[T.Zero-Shot]	-0.000	-0.018	0.018
ProblemType[limit]	0.011	-0.006	0.029
ProblemType[derivative]	0.011	-0.006	0.029
Difficulty[medium]	-0.011	-0.029	0.006
Difficulty[hard]	0.000	-0.018	0.018
Group Var	0.000		

C. Visualization of Model Coefficients

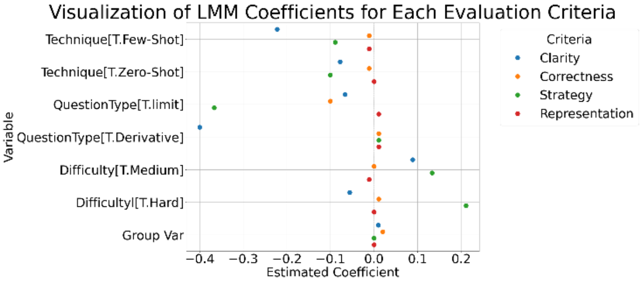

Figure 2 Coefficient Visualisation

The figure above presents a visualization of the coefficients from the Logistic Mixed Effects Model (LMM) for each predictor variable against the four evaluation criteria. Points located far from the vertical zero line indicate a stronger influence. The Figure 2 shows that Clarity and Strategy are the two criteria most affected by variations in prompting technique and problem type. Derivative problems have the largest negative impact on clarity, while limit problems have the largest negative impact on strategy. Conversely, an increase in problem difficulty shows a positive correlation with an improvement in strategy quality.

D. Result and Discussion

1) Result

The research findings consistently show that Chain-of-Thought (CoT) is the most superior prompting technique for generating high-quality calculus solutions, especially in the dimensions of clarity and strategy. This finding supports previous literature stating that a step-by-step thinking process facilitates the model in constructing more structured and logical solutions. Conversely, Few-Shot demonstrated a significant weakness in solution clarity, possibly due to a lack of explicit reasoning steps or errors in the provided examples.

The problem type also plays a crucial role. The model's excellent performance on Integral problems indicates that their stable and frequently occurring structure in the training data makes the model more 'proficient' at handling them. On the other hand, Limit problems proved to be the biggest challenge, significantly lowering the strategy and correctness scores, indicating the model's vulnerability to logical misconceptions in this topic.

Another interesting finding is the influence of problem difficulty. More difficult problems actually prompted the model to produce more strategic solutions, albeit at the expense of clarity. This suggests that ChatGPT responds more accurately to more complex challenges, especially when guided by appropriate prompts.

Overall, these results suggest that to maximize the effectiveness of ChatGPT in the context of mathematics education, the choice of prompting techniques should be made consciously and strategically, taking into account the type and complexity of the problem at hand.

2) Discussion

This research reinforces the fundamental findings of Wei et al. (2022) regarding the superiority of Chain-of-Thought (CoT) in eliciting complex reasoning[5]. However, it also adds a new nuance by demonstrating in line with the evaluation framework of Schorcht et al. (2024) that CoT is primary advantage lies in enhancing pedagogical values like clarity and strategy, not merely in the correctness of the final answer[1]. Furthermore, by focusing specifically on the calculus domain, this study moves beyond general mathematical analyses. Instead, it varies significantly by topic, as evidenced by the strong performance on integral problems compared to the significant difficulty with limit problems.

V. Conclusion

This research concludes that the prompting technique is a crucial factor that determines the quality of calculus solutions generated by Large Language Models (LLM). Specifically, the Chain-of-Thought (CoT) technique proved to be the most superior, not in significantly improving the correctness of the final answer, but rather in producing solutions with superior clarity and strategy. Furthermore, it was found that solution quality is highly dependent on the problem type, with limit problems consistently posing the greatest challenge for the model compared to integral problems.

The main implication of these findings is that for applications in educational contexts, CoT is the most recommended strategy to maximize the pedagogical value of AI. Considering the study's limitations of using only a single LLM model and a textbook-based dataset, future research can be expanded. Recommended future research directions include: (1) a comparative analysis across different LLM models to test the generalizability of the findings; (2) the use of more diverse problem types, such as word problems, to test contextual adaptability; and (3) evaluations involving students to directly measure the pedagogical benefit of the solutions. These steps will enrich the understanding of how to optimize the role of AI as an effective learning tool in the field of mathematics.

References

[1] S. Schorcht, N. Buchholtz, dan L. Baumanns, “Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques,” *Front. Educ.*, vol. 9, no. May, hal. 1–15, 2024, doi: 10.3389/feduc.2024.1386075.

[2] D. Wang, L. Xu, C. Cao, X. Fang, dan J. Lin, “A Systematic Review on Prompt Engineering in Large Language Models for K-12 STEM Education,” 2024.

[3] S. Imani dan L. Du, “MathPrompter: Mathematical Reasoning using Large Language Models,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 5, no. 1, hal. 37–42, 2023, doi: 10.18653/v1/2023.acl-industry.4.

[4] Y. Chen *et al.*, “Assessing the Impact of Prompting Methods on ChatGPT’s Mathematical Capabilities,” hal. 1–7, 2023, [Daring]. Tersedia pada: http://arxiv.org/abs/2312.15006

[5] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, hal. 1–43, 2022.

[6] U. Lee *et al.*, *Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education*, vol. 29, no. 9. Springer US, 2024. doi: 10.1007/s10639-023-12249-8.

[7] M. Nazari dan G. Saadi, “Developing effective prompts to improve communication with ChatGPT: a formula for higher education stakeholders,” *Discov. Educ.*, vol. 3, no. 1, 2024, doi: 10.1007/s44217-024-00122-w.

[8] Y. Wei, Y. Ma, T. P. Garcia, dan S. Sinha, “A consistent estimator for logistic mixed effect models,” *Can. J. Stat.*, vol. 47, no. 2, hal. 140–156, 2019, doi: 10.1002/cjs.11482.

[9] M. Muradoglu, J. R. Cimpian, dan A. Cimpian, “Mixed-Effects Models for Cognitive Development Researchers,” *J. Cogn. Dev.*, vol. 24, no. 3, hal. 307–340, 2023, doi: 10.1080/15248372.2023.2176856.

[10] S. Schulhoff *et al.*, “The Prompt Report: A Systematic Survey of Prompting Techniques,” 2024, [Daring]. Tersedia pada: http://arxiv.org/abs/2406.06608

[11] G. M. Engeseth *et al.*, “Mixed Effect Modeling of Dose and Linear Energy Transfer Correlations With Brain Image Changes After Intensity Modulated Proton Therapy for Skull Base Head and Neck Cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 111, no. 3, hal. 684–692, 2021, doi: 10.1016/j.ijrobp.2021.06.016.

[12] R. D. Dermawan dan Herdianto, “Meningkatkan Kinerja Output ChatGPT Melalui Teknik Prompt Engineering Yang Dapat Dikustomisasi,” *J. Soc. Sci. Res.*, vol. 4, no. 1, hal. 10646–10664, 2024.

[13] B. Wang *et al.*, “Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, hal. 2717–2739, 2023, doi: 10.18653/v1/2023.acl-long.153.

[14] Y. Li, “A Practical Survey on Zero-shot Prompt Design for In-context Learning,” *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, hal. 641–647, 2023, doi: 10.26615/978-954-452-092-2_069.

[15] D. Serhan dan N. Welcome, “Integrating ChatGPT in the Calculus Classroom: Student Perceptions,” *Int. J. Technol. Educ. Sci.*, vol. 8, no. 2, hal. 325–335, 2024, doi: 10.46328/ijtes.559.

[16] A. K. Iddrisu, I. Besing Karadaar, J. Gurah Junior, B. Ansu, dan D. A. Ernest, “Mixed effects logistic regression analysis of blood pressure among Ghanaians and associated risk factors,” *Sci. Rep.*, vol. 13, no. 1, hal. 1–13, 2023, doi: 10.1038/s41598-023-34478-0.

[17] E. J. Purcell, “Matematika A - Purcell, Calculus, 9th ed.” hal. 1–797, 2021.

[18] S. C. E. Fung, M. F. Wong, dan C. W. Tan, “Chain-of-Thoughts Prompting with Language Models for Accurate Math Problem-Solving,” in *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2023, hal. 1–5. doi: 10.1109/URTC60662.2023.10534945.

[19] M. G. Worku, A. B. Teshale, dan G. A. Tesema, “Determinants of under-five mortality in the high mortality regions of Ethiopia: mixed-effect logistic regression analysis,” *Arch. Public Heal.*, vol. 79, no. 1, hal. 1–9, 2021, doi: 10.1186/s13690-021-00578-4.

[20] P. Sterzinger dan I. Kosmidis, “Maximum softly-penalized likelihood for mixed effects logistic regression,” *Stat. Comput.*, vol. 33, no. 2, hal. 1–14, 2023, doi: 10.1007/s11222-023-10217-3.