MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

Free and paid large language model performance evaluation in answering academic technical questions

Ilyas Nur Rokhim [a,1*], Fandy Dwi Putra [a,2], Raditya Duta Pratama [a,3], Sayudha Hanif Saputra [a,4]

[a] Department of Informatics, STMIK AMIKOM Surakarta, Indonesia
[1] ilyas.130512@mhs.amikomsolo.ac.id, [2] fandy.130522@mhs.amikomsolo.ac.id, [3] raditya.130549@mhs.amikomsolo.ac.id, [4] sayudha.130527@mhs.amikomsolo.ac.id
* corresponding author

Article Info

Article History:

Received: July 29, 2026
Revised: August 6, 2026
Accepted: August 13, 2026

Keywords:
Artificial Intelligence, ChatGPT, GPT-4o, Large Language Models, Performance Evaluation

Abstract

The rapid advancement of Large Language Models (LLMs) has increased the adoption of artificial intelligence in higher education, particularly in supporting students with fundamental computer science tasks. Although ChatGPT is available in both Free (GPT-4o-mini) and Premium (GPT-4o) versions, empirical evidence comparing their performance in academic contexts remains limited. This study aims to compare the performance of both versions based on answer accuracy, information completeness, response clarity, and response time. A comparative experimental method was employed using ten computer science task scenarios covering programming, databases, and system visualization. The outputs were evaluated using a five-point Likert scale, while response time was measured quantitatively. The results indicate that both models achieved high answer accuracy. However, ChatGPT Premium produced more comprehensive, well-structured responses and demonstrated superior visualization capabilities. In contrast, ChatGPT Free delivered faster response times, making it more suitable for tasks requiring efficiency. These findings reveal a trade-off between response quality and processing speed. This study provides a comparative evaluation framework that can support students and educators in selecting the most appropriate ChatGPT service according to their academic needs.

1. Introduction

The development of Artificial Intelligence (AI) technology, especially Large Language Models (LLMs), has grown very rapidly in recent times and has changed the way users obtain, process, and automatically generate information [1]. LLM's proficiency in understanding the context of human language, generating text, writing program code, and supporting knowledge-based analysis makes it one of the widely adopted technologies in various fields, such as education, software development, information services, and the creative industries [2]. With the increase in access to this technology, many LLMs are now available to the public in the form of free and paid services, including ChatGPT, Gemini, Claude, and other similar platforms [3]. In higher education, the application of LLM is also increasingly widespread as a tool to support the learning process, especially in helping students understand programming concepts, databases, system analysis, and complete various academic tasks [4].

The increase in the use of LLM in the academic field has given rise to various perceptions about the quality of existing services. One of the views that is often held is that paid services always produce superior outputs to free services. However, this view has not been fully proven through empirical evaluation of real use by students [5]. In addition, the selection of AI services is not only influenced by the quality of the answers, but also considers the factors of cost efficiency, speed of response, and ease of use. Therefore, research is needed that can present an objective picture of the characteristics of each service so that users can choose the service that best suits their needs [6].

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

Various previous studies have assessed the performance of Large Language Models in a number of areas, such as logical reasoning, program code writing, mathematical problem solving, and evaluation of academic abilities [7], [8]. Most of the findings suggest that models with larger sizes tend to produce more complete answers and have better reasoning abilities. However, previous research has usually focused on general assessments of AI models or using standard benchmarks, so there is still rarely a direct comparison between the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o) services in the context of completing basic tasks that informatics students often face. In addition, aspects of visualization, completeness of information, and speed of response are still relatively rarely analyzed simultaneously in a single evaluation framework [9].

Based on the study, there are still several research gaps that need to be looked at more deeply. First, there is still little research that specifically compares Free ChatGPT services with Premium in the context of basic informatics tasks [10]. Second, most studies focus on the assessment of textual ability, while aspects such as visual representation, explanatory quality, and response speed have not been thoroughly analyzed. Third, there is still a rare study that utilizes testing scenarios that reflect students' academic activities as the basis for assessing the performance of the two services [11].

In response to the research gaps that have been identified, this study contributes in the form of a comparative evaluation framework to assess the performance of Free ChatGPT (GPT-4o-mini) versus ChatGPT Premium (GPT-4o) through ten basic informatics task scenarios that reflect students' academic activities. The assessment is carried out by considering four main indicators: the accuracy of the answers, the completeness of the information, the clarity of delivery, and the speed of response, so as to present a more comprehensive picture of the characteristics of each service. This approach is expected to enrich the evaluation study of Large Language Models at the higher education level and become a reference for users in choosing the AI service that best suits their academic needs [12].

Referring to the explanation above, this study aims to assess and compare the performance of the Free version of ChatGPT with ChatGPT Premium in completing basic informatics tasks. It is hoped that the research findings will not only present information about the difference in the quality of the output of the two services, but also become a reference for students and teachers in utilizing AI technology more effectively, efficiently, and in accordance with learning needs [13].

2. Research Methodology

This study applied a comparative experimental method with a descriptive quantitative approach to assess the performance difference between the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o) [14]. The choice of the experimental approach was based on its ability to test both models under similar conditions using the same set of test scenarios, so that differences in results could be observed objectively. Furthermore, a comparative approach is used to compare the output quality of each model based on predetermined evaluation indicators [15]. The assessment was carried out on four main dimensions: accuracy of answers, completeness of information, clarity of delivery, and speed of response. The first three indicators assess the quality of the answer content, while response speed measures the time efficiency it takes the model to produce output [16].

The stages of the research are shown in Figure 1, which consists of the preparation of test scenarios, the implementation of experiments, the assessment process, data analysis, and the drawing of conclusions.

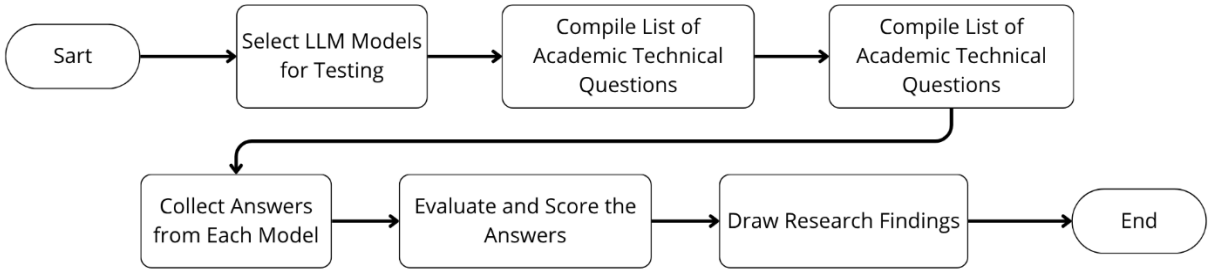

Fig. 1. Research Flow

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

2.1 Experimental Design

The experiment involved two ChatGPT services from the same provider: the Free version (GPT-4o-mini) and the Premium version (GPT-4o). Both models were chosen so that architectural differences between providers did not affect the results, so the comparison was focused on the variety of services offered to users. All tests were performed on identical devices, using the same browser, internet connection, and user account. Each scenario is run in a *new chat* to avoid the influence of the previous context on the model's response. Furthermore, all prompts are presented with no modifications or sample answers, ensuring each model is in the same test condition. To reduce the variation caused by network fluctuations or server load, each scenario is repeated three times, and the value used is the average of all three attempts.

2.2 Dataset Characteristics

The research dataset includes ten task scenarios that were purposively selected to reflect the academic activities of informatics students at the initial level. The selection of the ten scenarios took into account the representation of the material that most often appeared in the learning process, so this number was considered sufficient to provide a preliminary idea of the performance of the two models. The ten scenarios are divided into three groups, namely:

1. Programming and algorithms (4 scenarios)

Table 1. Programming Scenarios and algorithms

No	Skenario
1	Create a function to reverse words (e.g. "book" to "ukub") using Python.
2	Explain the difference between Stack and Queue data structures.
3	Correct errors (debugging) in Java code that mistyped mathematical formulas.
4	Create a simple program code to sort the numbers from small to large.

2. Database (3 scenarios)

Table 2. Database Scenarios

No	Skenario
1	Write the SQL SELECT command using the WHERE and ORDER BY conditions to bring up the student data.
2	Explain the difference between the INNER JOIN and LEFT JOIN commands in the database.
3	Create a simple database table (ERD) for the school library system.

3. System visualization (3 scenarios)

Table 3. System Visualization Scenarios

No	Skenario
1	Create a flowchart image that shows the flow of the user login process in an application.
2	Sketch a block diagram to illustrate the Client-Server architecture on the web.
3	Create an Entity-Relationship Diagram (ERD) visual image or chart for a sales system.

This grouping is intended to assess the model's ability on various types of tasks with different characteristics, ranging from logical reasoning, conceptual understanding, to the ability to generate visual representations.

2.3 Evaluation Instruments

The quality of the output was assessed by a rubric based on a five-point Likert Scale, with a value between 1–5. The evaluation instrument is formulated based on three main indicators that are commonly used to assess the quality of the output of the Large Language Model, namely:

1. Accuracy; which assesses the accuracy of the answer to the given problem.
2. Completeness; which assesses the completeness of the information and the depth of explanation.

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

3. Clarity; which assesses the level of readability, delivery structure, and ease of understanding of answers.

The assessment rubric was developed by adapting the AI output quality evaluation indicators from the previous research, then adjusted to the characteristics of the basic informatics tasks used in this study. Each answer is assessed using the same rubric so that all models are evaluated based on identical criteria.

Table 4. Research Variables

Variabel	Indicator	Scale
Accuracy of Answers	Accuracy of answers to questions	Likert 1-5
Completeness of Information	Completeness of materials and examples	Likert 1-5
Clarity of Delivery	Structure and ease of understanding	Likert 1-5
Response Speed	Time to answer complete (seconds)	Ratio

2.4 Response Time Measurement

To assess the answers from AI, the author uses the Likert Scale with values of 1 to 5. The assessment rubric is specially created to see the quality of the resulting text, program code, and image diagrams, such as those in Table 4.

Table 5. Testing Procedure

Score	Accuracy of Answers	Completeness of Information	Clarity & Ease of Understanding
5 (Very Good)	The answer is correct, the code can run without errors, and the diagram image fits according to the lecture material.	The explanation is complete, there are basic theories, code/visual examples, and alternative solutions.	The language is neat, the structure is structured, and uses a markdown format that is easy to read.
3 (Sufficient)	The basic logic is correct, but there is a slight typo in the code or the diagram image is a bit lacking.	Just tell me the main answer without any explanation of the supporting theory.	It's understandable, but the language is a bit stiff or twisted.
1 (Ugly)	The answer is completely wrong, or the AI is a delusion (data hallucination).	The answer is very short, cut, or not connected at all.	The language is messy and confusing for the reader.

The response time is defined as the interval between the prompt delivery and the completion of the model process resulting in the entire response. Measurements are carried out with a digital stopwatch on the same device to maintain consistency of observation. All experiments were run under stable internet network conditions and without any other activities that could affect system performance. The response time value in each scenario is taken from an average of several attempts, so that it can reduce variations caused by network conditions or service loads.

2.5 Bias Control

To strengthen the validity of the test results, a number of bias control measures were implemented during the testing process.

1. The instructions given to both models are in the form of exactly the same sentences.
2. Each test is run on a new conversation session so as not to be affected by the previous context.
3. All experiments used identical devices, browsers, accounts, and internet networks.

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

4. No Custom Instructions, Memory, or additional file uploads were used during the experiment.
5. The tests are conducted over a similar time frame to mitigate the impact of server load variations.
6. All answers were assessed with the same evaluation rubric so that each model was treated equally.

2.6 Data Analysis

Quantitative data is analyzed with descriptive statistics, which calculate the average of each evaluation indicator. After that, the test results of the two models were compared in each task category to identify performance differences in accuracy, completeness of information, clarity of delivery, and speed of response. The analysis was carried out comparatively, combining the results of the experiment with the characteristics of each model as well as the findings of previous studies, resulting in a more comprehensive interpretation of the advantages and limitations of the two services.

3. Results and Discussion

3.1 Quantitative Performance Test Results

Trials on ten task scenarios showed that both models produced high answer accuracy, with an average score of 5.0 across all test categories. These results show that both the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o) can complete basic informatics tasks such as simple programming, database concepts, and algorithm logic design with almost the same level of accuracy.

Table 6. Performance Test Results

Evaluation Dimensions	ChatGPT Free Version (Score 1-5)	ChatGPT Go/ Paid (Score 1-5)
Accuracy of Answers	5.0	5.0
Completeness of Information	4.2	4.6
Clarity & Easy to Understand	4.2	4.6
Response Speed (Time)	11.5 seconds	15 seconds

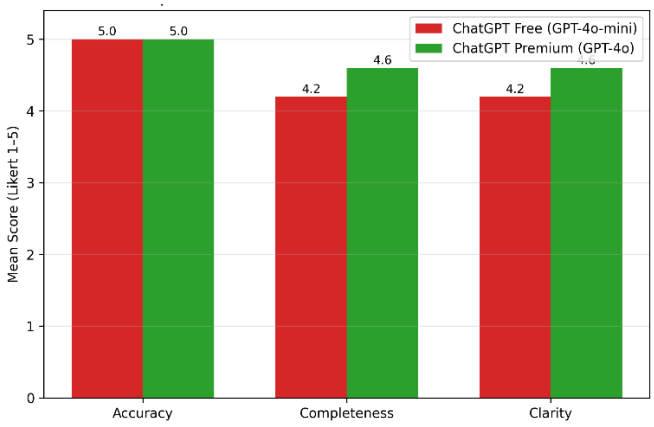

Fig. 2. Comparison of ChatGPT Free (GPT-4o-mini) and ChatGPT Premium (GPT-4o)

Figure 2 presents a comparison of the average scores of the two models on three evaluation dimensions, namely the accuracy of the answers, the completeness of the information, and the clarity of delivery. Based on the visualization, both models obtained the same accuracy score, which is 5.0, indicating that both have equal abilities in completing basic informatics tasks. However, differences are starting to be seen in the

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

aspects of completeness of information and clarity of delivery, where ChatGPT Premium obtained an average score of 4.6, while the Free version of ChatGPT obtained a score of 4.2.

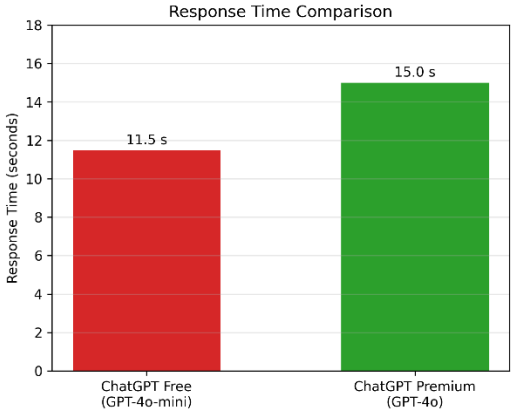

Fig. 3. Comparison of Response Time

Figure 3 shows the comparison of the average response time between the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o). The test results show that the Free version of ChatGPT produces faster responses with an average time of 11.5 seconds, while ChatGPT Premium takes about 15 seconds. This difference indicates that models with larger capacities require longer inference times because they produce more detailed, more structured, and take into account the broader context. Thus, there is a trade-off between output quality and response time efficiency.

3.2 Comparative Analysis by Task Category

To gain a more comprehensive understanding of the characteristics of each model, the test results were not only analyzed based on quantitative data, but also compared based on the category of tasks tested. This analysis aims to identify the advantages and limitations of the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o) on each type of academic assignment.

Table 7. Summary of Comparison of Results by Task Category

Category Task	ChatGPT Free	ChatGPT Premium	Superior models
Programming	True Code, Brief Explanation	Correct code, full explanation and better debugging	Premium
Basis Data	SQL True, Text-dominant ERD	SQL is true, ERD is more Structured	Premium
Visualization	Simple Diagram	More Complete and Informative Diagrams	Premium

Table 7 shows that the difference in performance between the Free version of ChatGPT (GPT-4o-mini) and ChatGPT Premium (GPT-4o) does not lie in the level of accuracy of the answers, but rather in the quality of the presentation of information in each category of tasks. In general, both models are able to generate correct answers in all test scenarios, but ChatGPT Premium consistently provides more complete, more structured, and more informative explanations than the Free version of ChatGPT. To provide a more detailed understanding of the characteristics of each model, the following discussion outlines the test results based on three categories of tasks, namely programming and algorithms, databases, and system visualization.

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (3) 2026, p. 119-127
E-ISSN : XXXX-XXXX

a. Programming and Algorithm Tasks :
In programming tasks, both models can generate precise code in terms of both syntax and basic logic, signifying that GPT-4o-mini and GPT-4o are already competent enough to handle basic-level programming problems. The difference arises in the debugging situation. ChatGPT Premium not only fixes syntax errors, but also explains the causes, offers alternative solutions, and suggests better code writing. In contrast, ChatGPT Free tends to provide shorter solutions with limited explanations. These findings suggest that the premium model is superior in the learning process because it is able to provide conceptual explanations, not just final answers.

b. Database Tasks
Within the database task group, both models manage to generate precise SQL commands and can adequately explain basic concepts such as the difference between INNER JOIN and LEFT JOIN. This shows that both models have sufficient knowledge representation capabilities in the field of databases. However, on the Entity Relationship Diagram (ERD) design task, ChatGPT Premium generates a more organized table structure with key attribute tagging and clearer relationships between entities. Whereas ChatGPT Free presents more explanations in the form of text, so it requires additional interpretation from the user. Thus, the premium model provides added value to tasks that require the presentation of information in a visual and structured manner.

c. Visual Representation Tasks
The most noticeable difference was observed in visualization work. The Premium version of ChatGPT can create visualizations in the form of flowcharts, client-server diagrams, and ERDs that are more informative and easy to understand, while facilitating the completion of academic assignments that require visual presentation of concepts. In contrast, the Free version of ChatGPT usually presents simple text-based or visual explanations. Although the data presented is conceptually correct, the visual quality remains below the standard of the premium model. These findings confirm that the multimodal function is the main advantage of the premium model, especially in learning activities that require visual media.

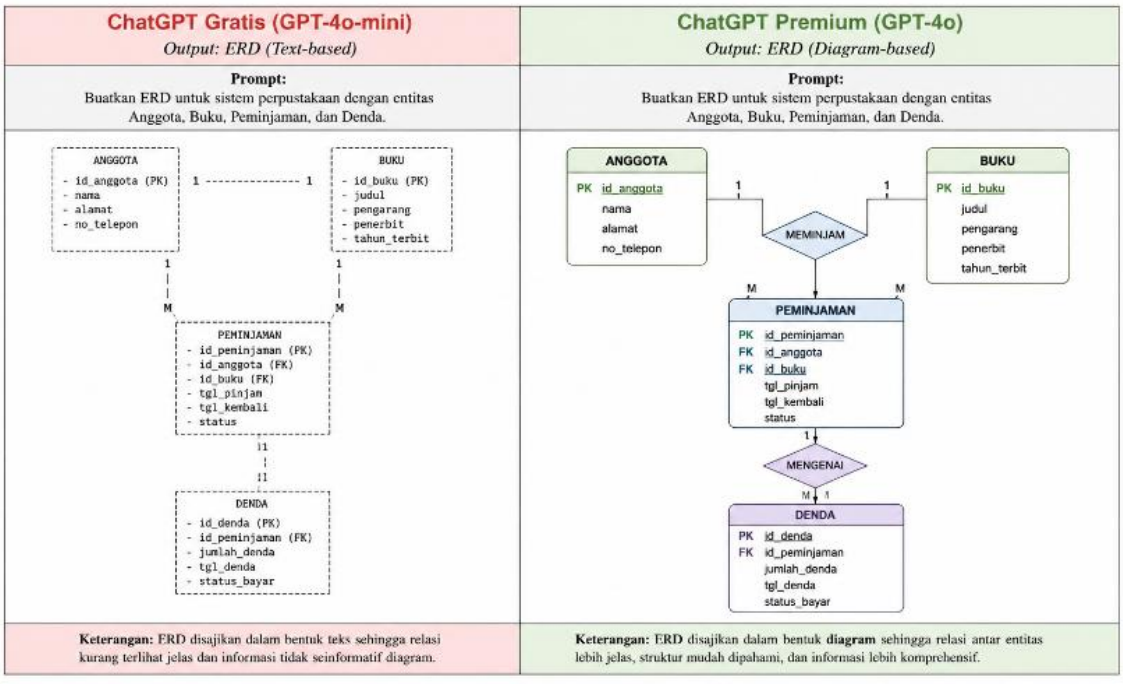

Fig. 4. Example of Comparison of the generated ERD

Figure 4 shows an example of the Entity Relationship Diagram (ERD) output generated by both models using the same prompt. The Free version of ChatGPT tends to generate text-based representations with relationships between tables that still require user interpretation. On the other hand, ChatGPT Premium is able to present the ERD in the form of a more structured diagram, showing entities, attributes, and relationships between entities more clearly. The visualization reinforces the results of the evaluation that multimodal capabilities are one of the main advantages of the premium model.

Based on the results of the analysis on the three task categories, it can be seen that improving the quality of output in premium services is not always followed by an increase in response speed. ChatGPT Premium produces more comprehensive, more structured, and better visualization capabilities, but it requires a slightly longer processing time than the Free version of ChatGPT.

The results of this study are consistent with a number of previous studies that show that larger language models tend to provide higher-quality answers to tasks that require reasoning, conceptual explanation, and multimodal skills. However, these quality improvements usually lead to longer inference times. Therefore, the selection of models must be adapted to the needs of the user. For college students, the free version of ChatGPT is enough to complete basic work such as searching for concepts, simple programming exercises, or writing SQL syntax. While the Premium version is more suitable for activities that require in-depth explanations, complex code debugging, diagramming, and visualization of learning concepts. Therefore, the decision to use a premium service should be based on the level of complexity of the task, not simply the assumption that paid services are always superior in all aspects.

4. Conclusion

This study aims to assess and compare the performance of the Free version of ChatGPT (GPT-4o-mini) with ChatGPT Premium (GPT-4o) in completing basic informatics tasks commonly encountered in college learning environments. The assessment is carried out with four main criteria, namely the accuracy of the answers, the completeness of the information, the clarity of delivery, and the speed of response, so that the comparison of the two services can be carried out objectively in the same usage scenario.

The results showed that both models displayed high accuracy in completing basic informatics tasks. However, ChatGPT Premium provides a more comprehensive, structured response, and is able to present more informative visualizations than the Free version. In contrast, the Free version shows a higher response speed, making it a better choice for time-efficient needs. These findings highlight the trade-off between output quality and response speed, so service selection should be tailored to the characteristics of the task to be performed.

The main contribution in this study lies in the design of a comparative evaluation framework that includes four dimensions: accuracy, completeness of information, clarity of presentation, and speed of response, to assess the two ChatGPT services in the context of basic informatics tasks. The findings of this study are expected to be a reference for students, lecturers, and educational institutions in utilizing Large Language Models (LLMs) more optimally and in accordance with learning needs. In addition, this work adds to the empirical evidence regarding the characteristics of generative AI in higher education settings. However, this study has a number of limitations. The test included only ten task scenarios representing the basic materials of informatics and only compared two models from a single AI service provider. Output quality assessment still relies on rubrics evaluated by humans, so it contains the potential for subjectivity. Therefore, follow-up studies should expand the number of scenarios, cover a wide range of disciplines, involve LLMs from other providers, and use more thorough evaluation methods, including inferential statistical analysis to test the significance of differences in performance between models.

References

[1] A. Alvin, R. Robet, and A. Tarigan, “Implementasi chatbot Otomatis Akademik Berbasis Web Menggunakan LLM dan Rule-Based System Studi Kasus : STMIK Time Implementation of Web-Based Academic Automated chatbot Using LLM and Rule-Based System Case Study : STMIK Time,” no. 3, pp. 651–665, 2025, doi: https://doi.org/10.26798/jiko.v9i3.2209.

[2] M. R. Kusuma and L. Ternado, “Evaluasi Kinerja Model LLM Terkuantisasi pada Tugas Bahasa Indonesia,” vol. 01, no. 01, pp. 25–38, 2025.

[3] C. K. Wigayha and U. Bunda, "Analisis Sistematis Penggunaan Large Language Models (Llms) Dan Artificial Intelligence (Ai) Untuk Peningkatan Literasi Digital Pada Jenjang Pendidikan Tinggi," vol. 1, no. 1, pp. 1–7, 2025.

[4] A. S. Susanto and L. Lisana, "Jurnal Locus : Penelitian & Pengabdian Chat AI LLM (Large Language Model) di Universitas : Tinjauan Sistematis terhadap Variabel Pendorong Adopsi , Kerangka Teoritis, dan Tren Global," vol. 4, no. 10, 2025. https://doi.org/10.58344/locus.v4i10.4324

[5] P. Studi, P. Jarak, J. Ilmu, U. P. Harapan, M. I. Komunikasi, and U. P. Harapan, "Penggunaan ChatGPT di Kalangan Mahasiswa dan Dosen Perguruan Tinggi Indonesia," vol. 14, no. 1, pp. 130–145, 2024. https://doi.org/10.35814/coverage.v14i2.6058

[6] A. Yuanda *et al.*, "Analisis Pemanfaatan Fitur Gratis Dan Premium Chatgpt Sebagai Materi Presentasi Mahasiswa Uinsu," vol. 4, no. April, pp. 36–47, 2026.

[7] A. N. Annas, G. Wijayanto, D. Cahyono, and M. Safar, "Pelatihan Teknis Penggunaan Aplikasi Artificial Intelligences (AI) Chat Gpt Dan Bard AI Sebagai Alat Bantu Bagi Mahasiswa Dalam Mengerjakan Tugas Perkuliahan," vol. 4, no. 1, pp. 332–340, 2024.

[8] M. B. Ibrahim, I. Sari, and L. Lubis, "Optimalisasi Riset Berbasis Artificial Intelligence dalam Meminimalisir Prokrastinasi Akademik Mahasiswa," vol. 5, no. 2, pp. 100–112, 2025, doi: https://doi.org/10.59395/jitp.v5i2.127.

[9] S. Yin *et al.*, "A survey on multimodal large language models," no. October, 2024, doi: https://doi.org/10.1093/nsr/nwae403.

[10] K. Buatan, P. Personal, and S. D. Pembelajaran, "Education Journal : Journal Education Research and Development," no. November 2022, pp. 158–166.

[11] S. Berbasis, K. Tpack, and D. A. N. Tam, "Generative ai," vol. 11, 2026.

[12] A. J. Nabila, W. Pebrianti, and H. Setiawan, "Determinants of Continuance Intention Toward ChatGPT Premium Among Indonesian University Students : An Information Systems Success Model Perspective for Higher Education Management," vol. 05, no. 02, pp. 1409–1427, 2026. https://doi.org/10.61987/jemr.v5i2.1892

[13] N. Luh, C. Pendiasari, F. D. Acarya, U. Hindu, N. I. Gusti, and B. Sugriwa, "Pengaruh Artificial Intelligence (AI) terhadap Efektivitas Proses Pembelajaran di Era Digital," vol. 4, no. 1, pp. 1–11, 2026. https://doi.org/10.61292/cognoscere.302

[14] D. A. N. E. Program, "Jurnal Riset Ilmiah," vol. 3, no. 1, pp. 67–76, 2026. https://doi.org/10.62335/sinergi.v3i1.2228

[15] I. Yunita, D. Rahmadiyanti, A. Aulia, N. Intan, D. Pandini, and I. Fauzi, "JPK : Jurnal Pendidikan dan Kebudayaan. Analisis Komparatif Laporan Penelitian Tindakan Kelas dan Penelitian JPK : Jurnal Pendidikan dan Kebudayaan," vol. 03, no. 02, pp. 103–114, 2026. https://doi.org/10.56842/jpk.v3i2.1051

[16] A. Manan and U. Nasri, "Tantangan dan Peluang Pendidikan Bahasa Arab : Perspektif Global," vol. 9, pp. 256–265, 2024. https://doi.org/10.29303/jipp.v9i1.2042