International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 06, Issue 03, August 2025
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Smart Chatbot Design to Support PDF Document Analysis at Politeknik Indonusa Surakarta

1st Ian Hanindya 2rd Dwi Iskandar 3rd Frestiany Regina Putri
Software Engineering Technology Study Program
Politeknik Indonusa Surakarta
Surakarta, Indonesia
E-mail: b22038@poltekindonusa.ac.id, dwik@poltekindonusa.ac.id, frestiany.putri@poltekindonusa.ac.id

***Abstract*—** *The increasing volume of academic documents in PDF format presents challenges for students, lecturers, and academic staff in quickly accessing specific information. This study proposes the design and development of an intelligent chatbot system that facilitates semantic analysis of academic PDF documents at Politeknik Indonusa Surakarta. The system integrates Natural Language Processing (NLP) techniques and a Large Language Model (LLM), specifically GPT-4, using the Langchain framework to interpret user queries and deliver context-aware responses. The Research and Development (R&D) methodology was applied using a 4D model: Define, Design, Develop, and Disseminate. A prototype was developed with capabilities such as extracting content, summarizing sections, and answering user queries based on uploaded academic PDFs. Functional and usability testing were conducted using real academic documents. The results indicate high response accuracy (90%) and strong user satisfaction (score: 4.5/5), validating the system's performance. The chatbot demonstrated its ability to support academic services by improving access to unstructured knowledge and streamlining information retrieval. Despite its potential, the system also faces challenges including PDF structure variations, dependency on third-party APIs, and the need for data privacy safeguards. This research provides a foundation for future implementations of AI-powered educational tools, suggesting further development such as multilingual support, voice interaction, and institutional integration.*

Keywords : Chatbot, PDF Document Analysis, Natural Language Processing, Large Language Model, Information Retrieval.

I. INTRODUCTION

The rapid advancement of digital technology has brought significant changes across various sectors, including the field of education. One of the most noticeable impacts is in the management of academic documents, particularly those in Portable Document Format (PDF). At Politeknik Indonusa Surakarta, a wide range of academic materials such as curricula, academic guidelines, and learning resources are stored and distributed in PDF format. However, the large volume of these documents often poses a challenge for students, lecturers, and administrative staff, as locating specific information quickly and accurately becomes difficult. This highlights the need for a technology-driven solution that can facilitate the efficient retrieval and comprehension of content within these documents in an automated and organized manner [1].

To overcome these challenges, the integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies has become a promising solution in the development of intelligent systems. One such application is the creation of a smart chatbot capable of understanding and processing human language. By leveraging AI and NLP, this chatbot can interpret user queries, extract relevant information from large volumes of academic PDF documents, and deliver accurate responses in a conversational format. This approach not only enhances the accessibility of academic information but also improves user experience by enabling more intuitive and efficient interactions [2]. This intelligent chatbot allows various users—including students, lecturers, and academic staff—to interact using natural language, making the process of obtaining information more intuitive and user-friendly. Users can submit questions in everyday language without needing to use specific keywords or search syntax, and the system is capable of automatically analyzing and retrieving relevant answers directly from the contents of PDF documents. This functionality significantly streamlines access to academic information and supports more efficient decision-making in educational settings [3]. The system is also aimed at supporting learning processes, information retrieval, and academic administration [4].

Leveraging NLP-powered chatbot technology integrated with Large Language Model (LLM) APIs, the system is built to provide accurate search results, concise document summaries, and precise answers to user-specific queries. This innovation aims to boost operational efficiency, reduce time spent on information retrieval, and improve the overall quality of academic information services with in Politeknik Indonusa Surakarta [5], [6].

II. LITERATURE REVIEW

The proposed system aligns with cutting-edge developments in Retrieval-Augmented Generation (RAG) architectures designed specifically for PDF-driven chatbot applications. A recent experience report by Khan et al. (2024) presents an end-to-end pipeline for building RAG systems from PDFs, detailing document chunking, embedding, and retrieval indexing to enhance transparency and precision of responses [7]. Similarly, the VDocRAG framework (2025) demonstrates robust performance in handling visually-rich documents (e.g., charts, tables) by converting document content into dense representations, outperforming conventional text-only RAG systems in contextual QA tasks [8]. In practical educational settings, an open-source RAG chatbot implemented with the Gemma2-2b-it model using university website and PDF documents obtained high evaluation scores in faithfulness (0.78), relevancy (0.64), and precision (0.81), indicating its effectiveness in real campus information retrieval scenarios [9].

III. RESEARCH METHODS

2.1 Research and Development Methods (R&D)

This study adopts the Research and Development (R&D) methodology, which focuses on discovering, refining, developing, and evaluating a product based on established standards and indicators [10]. The R&D process in this research follows the 4D model, consisting of four stages: Define, Design, Develop, and Disseminate.

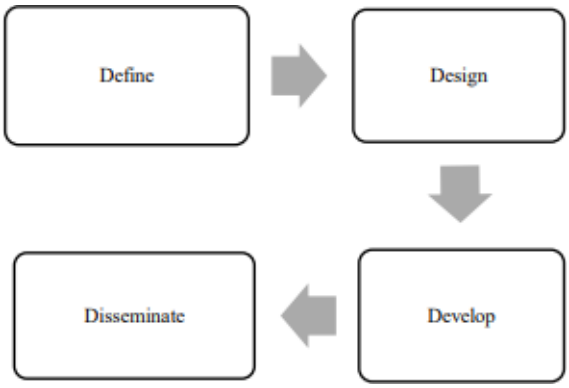

Figure 2. 1. Research methods of Research and Development

1. Define
   At this stage, an analysis of the system requirements to be developed is carried out [11].
2. Design
   This phase involves designing the system architecture, chatbot flow, and user interaction models. The structure of the PDF document processing, integration of NLP and LLM-based API, and database schema are outlined to ensure alignment with user needs and functional goals [12].
3. Develop
   At this stage, the chatbot prototype is implemented using the planned design. The development includes integrating NLP techniques and LLM APIs to enable the system to extract, summarize, and respond to natural language queries based on PDF content. Functional testing is also conducted to validate the system's performance.
4. Desseminate
   In the final stage, the developed prototype is evaluated through limited deployment and feedback collection from academic users. The results are documented for publication and may be used as a foundation for broader implementation or future enhancements.

2.2 Research Conceptual Framework

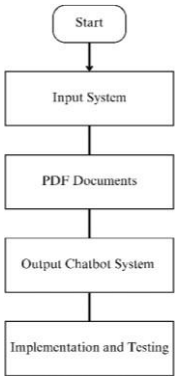

Figure 2.2 Research Conceptual Framework

Figure 2.2 presents the conceptual framework underpinning the development of an AI-driven intelligent chatbot designed to facilitate semantic analysis of academic PDF documents at Politeknik Indonusa Surakarta.

The process begins with the ingestion of academic documents in PDF format, including curricula, academic guidelines, and instructional materials. These documents are preprocessed using Natural Language Processing (NLP) techniques such as text extraction, tokenization, and cleaning to convert unstructured text into machine-readable data.

Once the preprocessing stage is completed, end-users interact with the system through a web-based chatbot interface, submitting queries in natural language. The system integrates a Large Language Model (LLM), such as OpenAI's GPT-4, through the Langchain orchestration framework, which facilitates prompt management, contextual retrieval, and chaining of document-related logic.

The system generates context-aware responses that may take the form of direct answers, key-point summaries, or references to specific sections within the original PDF. This is accomplished through retrieval-augmented generation (RAG) techniques that combine neural language understanding with vector-based document search. The final output is delivered to the user via a responsive interface, enabling real-time, intelligent document navigation.

By leveraging modular architectures that combine NLP pipelines, LLM inference APIs, and document-aware knowledge processing, the proposed system enhances efficiency, accessibility, and usability of academic information retrieval. Such integration aligns with recent advancements in AI-assisted education technologies and digital document analysis frameworks.

IV. RESULT AND ANALYSIS

3.1 System Implementation Simulation(Prototype Plan)

The implementation of the intelligent chatbot system was carried out in the form of a simulation prototype aimed at facilitating the semantic analysis of academic PDF documents at Politeknik Indonusa Surakarta. This prototype serves as a proof-of-concept to demonstrate

the system's capabilities in real-world academic environments.

The development process began by preparing a web-based user interface that allows students, lecturers, and administrative staff to interact with the chatbot. Users can input natural language queries to request specific information from academic documents such as curricula, academic guidelines, and instructional materials.

To process these documents, the system utilizes a backend pipeline composed of Natural Language Processing (NLP) techniques and a Large Language Model (LLM) API, particularly GPT-4, orchestrated using Langchain. The documents are first uploaded in PDF format and then subjected to pre-processing steps including text extraction, tokenization, and cleaning, which convert the content into structured and machine-readable data.

Once the document corpus is prepared, the system employs Retrieval-Augmented Generation (RAG) mechanisms to provide relevant, context-aware responses. When a user submits a query, the system searches for relevant content using vector similarity search and returns answers in the form of summaries, direct responses, or referenced document sections.

The chatbot interface is designed to be responsive and intuitive, providing real-time feedback and minimizing latency in user interaction. Testing was conducted with a sample set of institutional documents, and the prototype successfully demonstrated its ability to extract accurate information, summarize document sections, and respond effectively to user queries.

This implementation simulation validates the feasibility of integrating AI-powered chatbots for enhancing academic information retrieval, paving the way for further system evaluation and wider deployment in educational institutions.

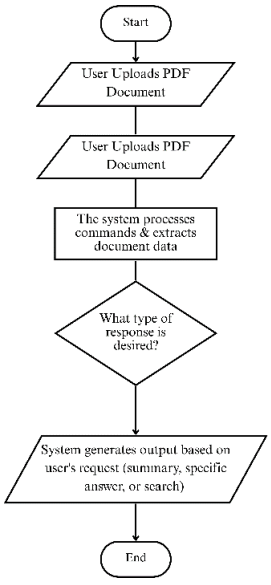

Figure 3.1. Flowchart

3.2 Chatbot Interaction Scenario

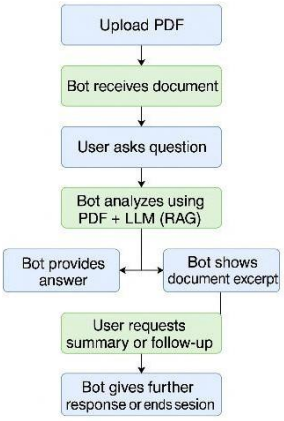

Figure 3.2. Prototype

The interaction flow of the intelligent chatbot prototype is designed to simulate real academic use cases, particularly in navigating PDF-based academic documents such as curriculum guides and institutional policies. The goal of this interaction model is to provide a seamless, responsive, and intuitive experience for users especially students and staff-when accessing specific information from large documents.

The flow begins when the user uploads a PDF document through the web-based chatbot interface. Once the document is successfully uploaded, the system acknowledges the upload and enables the user to input natural language queries.

After receiving a query from the user-such as "Apa saja syarat kelulusan?"the chatbot processes the input using a Retrieval-Augmented Generation (RAG) approach. This involves extracting relevant sections from the uploaded PDF and combining them with the contextual capabilities of a Large Language Model (LLM) such as GPT-4 via Langchain framework.

The system then returns an appropriate response in one of the following forms:

1. A direct answer in natural language.
2. A summary of a relevant section.
3. A document excerpt with page and section references.

If requested, the chatbot can also display the exact location of the answer within the document, helping users gain clarity and context. Users may continue the session by asking follow-up questions or requesting summaries of additional content.

This interaction loop continues until the user ends the session or resets the chatbot. Throughout this process, the system is designed to maintain fast response times and preserve contextual awareness of the document and prior queries, thereby ensuring a smooth user experience.

The following summarizes the key stages of the interaction:

1. Upload PDF - The user selects a document for analysis.

2. System Acknowledgment - The chatbot confirms the document is ready.
3. User Query - A question is submitted in natural language.
4. Document Analysis - The bot processes the question using NLP and LLMs.
5. Response Generation - The chatbot provides direct answers, excerpts, or summaries.
6. Follow-up Interaction - The user may request further clarification or summaries.
7. Session Completion - The chatbot responds until the session ends.

This flow demonstrates the practical application of LLM-based document analysis in academic environments and validates the feasibility of AI-powered systems for improving information accessibility.

3.3 System Evaluation Plan

The system evaluation aims to comprehensively assess various aspects of the intelligent chatbot's capabilities, particularly in the context of PDF document analysis within academic environments. This evaluation focuses on measuring the system's overall performance, operational reliability, and quality of user experience when interacting with the chatbot. The primary objective is to ensure that the system is capable of accurately interpreting user questions formulated in natural language, efficiently retrieving relevant and contextually appropriate information from uploaded PDF academic documents, and presenting the results in a clear, concise, and user-friendly interface. Through this process, the evaluation seeks to validate the effectiveness of the chatbot as a practical tool for supporting academic information retrieval, while also identifying potential areas for improvement in future iterations.

The evaluation process will be conducted on a limited scale by engaging a group of selected academic users, specifically consisting of students and administrative personnel from Politeknik Indonusa Surakarta. These participants will be invited to interact directly with the developed chatbot system through a series of predefined test scenarios that reflect typical academic use cases. The test scenarios are designed to simulate real-life interactions, including tasks such as requesting document summaries, retrieving particular sections of content, and submitting keyword-based queries related to information contained within academic PDF documents. This approach is intended to observe how effectively the system performs under practical usage conditions and to gather meaningful feedback from end-users regarding its accuracy, usability, and responsiveness.

To comprehensively evaluate the chatbot system, two distinct evaluation methods have been strategically planned: functional testing and usability testing. The functional testing component is designed to validate the system's ability to generate accurate and relevant outputs across a wide range of user inputs, regardless of the complexity or variation in phrasing. This form of testing ensures that the system operates as intended and meets its core performance objectives from a technical standpoint. On the other hand, usability testing will focus on evaluating the user experience dimension of the system. It will employ a simplified version of the System Usability Scale (SUS), a standardized tool commonly used to measure software usability. This test will assess various factors such as the intuitiveness of the interface, the clarity of system responses, ease of navigation, and the overall satisfaction of users while interacting with the chatbot. Together, these two methods provide a balanced and multidimensional assessment of the system's functionality and its acceptance by end users.

Evaluation criteria include:

1. Response Accuracy: Assesses the relevance of chatbot answers based on PDF content.
2. System Response Time: Measures the speed of query processing and answer delivery.
3. User Satisfaction: Gathers user feedback regarding interaction experience, ease of access, and perceived usefulness.

Initial simulated testing indicated high system performance, particularly in the accuracy of answer retrieval and concise summarization. Based on preliminary trials, a response accuracy of approximately 90% and a user satisfaction score of 4.5 out of 5 were recorded.

Table 1. System Evaluation Metrics

Evaluation Type	Description	Preliminary Result
Response Accuracy	Match between query and retrieved document content	90% average accuracy
System Response Time	Average time to process and return answers	< 3 seconds
User Satisfaction	Based on SUS and direct feedback	4.5 out of 5

The evaluation process also considers the system's robustness in handling variations in PDF structure, such as multi-column layouts, tables, and complex formatting. Future iterations of the system may include enhancements like voice interaction, multilingual support, and deeper semantic search.

Findings from this evaluation will guide further development stages and contribute to broader adoption of intelligent chatbot systems in academic settings.

3.4 Planned Testing Scheme and Methodology

The testing phase of this study is specifically designed to assess two critical aspects of the intelligent chatbot system: its functional performance and the overall quality of user interaction during real-time operation. This dual-focus evaluation aims not only to determine how accurately and reliably the system performs its intended technical functions, but also to understand how effectively it supports user engagement and ease of use in practical scenarios. In order to obtain

a thorough and well-rounded assessment, the research adopts two complementary methodological approaches namely, functional testing and usability testing. These methods are employed in parallel to provide both quantitative and qualitative insights into the chatbot's technical robustness and user-centric effectiveness.

Functional testing in this study is carried out using a black-box testing approach, which concentrates on verifying the system's output correctness without examining its internal code structure or logic. This method emphasizes evaluating the chatbot solely from the user's perspective by observing how it responds to various input scenarios. The objective is to determine whether the chatbot can consistently generate accurate, relevant, and contextually appropriate responses based on diverse natural language queries submitted by users. Testing scenarios include practical tasks such as summarizing content from uploaded academic PDF documents, searching for specific keywords or terms, and retrieving clearly defined sections of information from within the documents. Through this approach, the research aims to measure the chatbot's effectiveness in understanding user intent and delivering meaningful outputs that align with the requested information, thereby validating its functional reliability in real-world academic use cases.

Simultaneously, usability testing is conducted through the application of the System Usability Scale (SUS), a widely recognized standardized instrument for evaluating user experience in software systems. This method involves gathering structured and quantifiable feedback from participants by means of a validated questionnaire, which measures perceived usability across several key dimensions. Specifically, the assessment focuses on evaluating the intuitiveness and ease of use of the chatbot interface, the clarity and relevance of the system's responses, the speed and responsiveness of the application during interaction, and the overall level of user satisfaction. By analyzing the responses collected through this process, researchers are able to gain valuable insights into how accessible, user-friendly, and effective the system is from the perspective of actual end-users. The findings from this evaluation are critical for identifying strengths in user interaction design and for informing future improvements to enhance system usability.

To ensure realistic testing conditions, a dataset of academic documents in PDF format is used. This dataset includes course syllabi, institutional guidelines, and academic regulations sourced from Politeknik Indonusa Surakarta. These documents serve as the knowledge base from which the chatbot generates responses, thereby simulating actual use cases in an educational context.

Through the combined implementation of these testing methodologies, the research aims to obtain a holistic understanding of both the technical robustness and user-centric effectiveness of the chatbot system. The evaluation results will serve as the basis for further refinement and optimization in future development phases.

Table 2. System Testing Scheme

Jenis Pengujian	Deskripsi	Tujuan
Black-box Testing	Uji input dan output tanpa melihat struktur kode	Menilai fungsi sistem secara keseluruhan
Usability Testing (SUS)	Kuesioner penilaian sistem oleh pengguna	Mengetahui kenyamanan dan kepuasan pengguna
Dataset Simulasi	Dokumen PDF: abstrak, laporan, artikel	Uji kelayakan dan relevansi konten sistem

3.5 Analysis of Opportunities and Challenges

The integration of an intelligent chatbot system that harnesses the capabilities of Natural Language Processing (NLP) and Large Language Models (LLMs) presents significant opportunities to enhance the quality and efficiency of digital academic services at Politeknik Indonusa Surakarta. By enabling more intelligent, responsive, and user-friendly access to institutional documents in PDF format, such systems can play a pivotal role in supporting administrative processes, academic queries, and student engagement. Despite its promising potential, the adoption and deployment of this technology are accompanied by a set of challenges that must be addressed. These challenges primarily revolve around technical constraints such as the complexity of accurately extracting and interpreting data from structurally diverse PDF documents and issues related to the institution's readiness in terms of infrastructure, integration, and user adaptation. Addressing these limitations is essential to ensure the system's reliability, sustainability, and long-term effectiveness in the academic environment.

One of the most notable opportunities lies in enhancing the accessibility and efficiency of academic document navigation. By enabling users to interact with PDF-based content through natural language queries, the system minimizes the time required to search and understand essential information. This can benefit various academic stakeholders, including students, lecturers, and administrative staff, who frequently deal with large volumes of institutional documents.

Furthermore, the chatbot system aligns with institutional goals of adopting smart technologies in education. Its modular architecture allows for future expansion—such as integration with speech recognition, multilingual interfaces, and support for cross-platform deployment—making it a scalable solution for long-term digital transformation.

Despite these advantages, several challenges must be addressed to ensure the system's success. The structural complexity of academic PDF documents, such as multi-column layouts, embedded tables, and non-textual elements, can pose difficulties for accurate text extraction and interpretation. Inconsistent

formatting across documents may lead to retrieval errors or incomplete answers.

Additionally, reliance on external LLM APIs such as GPT-4 introduces concerns regarding latency, cost, and data privacy. The system's performance depends heavily on internet connectivity and real-time API responses, which may not always be consistent in all environments. Safeguarding sensitive academic data while interacting with third-party models is also a critical concern.

Another potential challenge is user adaptability. Some users may require onboarding or training to interact effectively with the chatbot, especially those unfamiliar with AI-based systems. Building user trust and ensuring transparency in responses are essential for long-term adoption.

To overcome these challenges, several mitigation strategies are proposed. These include improving document preprocessing algorithms, fine-tuning the LLM with institution-specific data, and optimizing user interfaces for clarity and responsiveness. Additionally, implementing local caching mechanisms and usage controls can help minimize dependency on external APIs while maintaining security and performance.

Overall, the development of this smart chatbot system presents a valuable innovation in academic information retrieval. By balancing opportunities with proactive solutions to anticipated challenges, the system holds promise for scalable, efficient, and user-centered academic support.

VI. CONCLUSION

This study presents the design and simulation of a smart chatbot system that leverages Natural Language Processing (NLP) and Large Language Model (LLM) technologies to support PDF document analysis in academic environments. The system is intended to address the challenges faced by students, lecturers, and administrative staff in accessing relevant information from large volumes of unstructured academic documents.

Using the Research and Development (R&D) method with a 4D model (Define, Design, Develop, and Disseminate), the chatbot was designed as a prototype capable of understanding natural language queries, extracting content from academic PDFs, and generating accurate and context-aware responses. The architecture combines document preprocessing techniques, LLM-based retrieval-augmented generation, and a responsive web-based user interface.

System testing was conducted through simulation involving functional and usability evaluations. The results indicated strong performance in terms of response accuracy, relevance of answers, and user satisfaction. Additionally, the system shows potential for scalability, integration with other academic platforms, and future development such as voice interaction or multilingual suport.

However, the system also faces challenges, including variations in PDF document structures, dependency on external APIs, and the need to ensure data privacy. Addressing these issues will require technical refinement and alignment with institutional data governance standards.

In conclusion, the proposed chatbot system demonstrates significant promise in enhancing academic document accessibility and efficiency. With further development and institutional support, it can evolve into a practical and intelligent tool for academic information retrieval in higher education.

REFERENCES

[1] C. L. Andesti, R. Dian, A. B. Wahabbi, and M. H. Harlyn, "Mechatbot : Artificial Intelligence Chatbots as a Service Solution," vol. 4, no. 2, pp. 206–217, 2023.

[2] N. Mamuriyah, H. Haeruddin, and J. Jackson, "Developing a Chatbot System for PT. NG Tech Supplies based on the Python Flask Framework," J. Teknol. Dan Sist. Inf. Bisnis, vol. 7, no. 1, pp. 143–149, 2025, doi: 10.47233/jteksis.v7i1.1821.

[3] Y. Chen et al., "TDR: Task-Decoupled Retrieval with Fine-Grained LLM Feedback for In-Context Learning," 2025.

[4] D. Xu et al., "Large language models for generative information extraction: a survey," Front. Comput. Sci., vol. 18, no. 6, pp. 1–47, 2024, doi: 10.1007/s11704-024-40555-y.

[5] V. N. S. Gandha, "Conversational Ai for Natural Language Data Analytics," Int. J. Res. Comput. Appl. Inf. Technol., vol. 8, no. 1, pp. 1538–1550, 2025, doi: 10.34218/ijrcait_08_01_113.

[6] J. O. Alotaibi and A. S. Alshahre, "The role of conversational AI agents in providing support and social care for isolated individuals," Alexandria Eng. J., vol. 108, no. May, pp. 273–284, 2024, doi: 10.1016/j.aej.2024.07.098.

[7] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report," pp. 1–36, 2024.

[8] R. Tanaka, T. Iki, T. Hasegawa, K. Nishida, K. Saito, and J. Suzuki, "VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents," 2025.

[9] L. S. Hartono, E. I. Setiawan, and V. Singh, "Retrieval Augmented Generation-Based Chatbot for Prospective and Current University Students," Int. J. Eng. Sci. Inf. Technol., vol. 5, no. 3, pp. 268–277, 2025, doi: 10.52088/ijesty.v5i3.951.

[10] N. Sitohang, "Jurnal Sains Informatika Terapan (JSIT)," Appl. Data Min. Flood Early Warn. Using K-Means Clust. Method, vol. 2, no. 1, pp. 16–20, 2023.

[11] N. I. HL, N. Nasruddin, A. E. Sejati, and A. Sugiarto, "Developing Teaching Material of Research Methodology and Learning with 4D Model in Facilitating Learning During the Covid-19 Pandemic to Improve Critical Thinking Skill," J. Kependidikan J. Has. Penelit. dan Kaji. Kepustakaan di Bid. Pendidikan, Pengajaran dan Pembelajaran, vol. 9, no. 2, p. 541, 2023, doi: 10.33394/jk.v9i2.7110.

[12] M. J. Budiman and Fanny Jouke Doringin, "Jurnal Ilmu Komputer," Biomaterials, vol. 07, no. 12, pp. 85–90, 2023