Development of a Chatbot Education System for Protected Marine Animals in Papua Using the Large Language Models Method

Anang Alfikran[1], Anggun Maimunah[2], Annisa Iriani Tawainella[3], Suhardi Aras[4]
[1,2,3]Informatics Engineering Studies Program, Muhammadiyah Sorong University, Indonesia.

ABSTRACT

The research aims to raise awareness about protected marine species in Papua through a chatbot providing accessible information. This strengthens public and student understanding of marine conservation. Methods include problem analysis, literature study, data collection, and system analysis and design using PyMuPDF and a Large Language Model (LLM). Data from scientific journals, books, e-books, and websites is processed through chunking and embedding to create vector representations. The FITZ library stores these vectors, enabling the chatbot to find similarities and provide relevant answers. Results show the chatbot accurately delivers information on Papua's protected marine animals. Testing with BERTScore indicated a high semantic correlation between chatbot responses and reference data. User satisfaction surveys reveal positive contributions to understanding marine conservation. The chatbot is relevant and satisfying, though performance and functionality could be improved with advanced technology. Further research should enhance data quality and answer consistency to better meet user needs.

Keywords: Chatbot, Protected Sea Animals, Papua, Large Language Models, BERTScore.

Corresponding Author:
Suhardi Aras
Informatics Engineering Studies Program
Muhammadiyah Sorong University
Jl. Pendidikan No 27 Sorong, 98416, Indonesia.
Email : suhardi.aras@um-sorong.ac.id

Article history:
Received May 1, 2024
Revised Jul 20, 2024
Accepted Aug 11, 2024

1. INTRODUCTION

Indonesia is a country with abundant natural resources (SDAs). One of the wealth of Indonesia's SDAs is in the marine resources sector. Indonesia has about 60,000 square kilometres of coral reefs. The area of this reef covers 51% of the total coral reservoir in Southeast Asia and 18% of the world's total. In this area, there is a very high coral diversity with more than 500 hard coral species that can be found (Ramadhan, Helena, and Nurrahman 2024).

Papua is one of the provinces of Indonesia located on the east end of Nusantara. The waters of Papua are part of the world's coral reef triangle known as the Coral Triangle which is home to thousands of marine species. Among the many species, there are some marine animals whose status is protected because of their endangered habitat due to various factors, such as illegal hunting, habitat degradation, climate change, and other human activities. Protected marine animals in Papua are varied, including: blue whales, buckwales, sperm whals, bottle-nosed dolphins, dugongs, green turtles, squirrels, frogs, manta fish, whale sharks, sea snails, and kakatua fish (Bawole and Megawanto 2017).

The high degree of diversity in coral reef ecosystems provides important ecological functions as shelters, breeds, reproduces, and seeks food for a variety of species. It's a major factor in increasing fish biomass within the ecosystem. The healthier the coral reef, the more space available for the fish to continue its life cycle. The waters in Papua are the basis for studies to assess the condition of protected marine animal ecosystems, useful to add information about protected species in the region (Prasetya et al. 2014).

The research is aimed at raising public awareness and knowledge about protected marine species. By providing easily accessible and understandable information, the chatbot is expected to strengthen public and student understanding of the importance of conserving marine animals and their ecosystems. The chatbot aims to reduce information gaps and provide accurate information based on the latest data and knowledge. In addition, these chatbots allow users to learn independently by providing access to educational resources anytime and anywhere without the need for direct help from educators.The use of modern technology also demonstrates practical applications of artificial intelligence in environmental education and conservation, as well as collecting user interaction data for research and evaluation, in order to improve the effectiveness of chatbots. By supporting education programmes and government policies, the system also contributes to joint efforts to protect the marine biodiversity of Papua (Aditya and Al-Fatih 2017).

2. RESEARCH METHODS

This research is a systematic and thorough process of digging and investigating specific problems using scientific methods. The objective is to collect, process, analyze data carefully, and draw conclusions in a structured and objective manner. All these steps are taken to solve problems or test hypotheses in order to acquire knowledge that is useful to human life. The stages of the research used in completing the research patch in Figure 1 below (Informatika and Abdurrab, n.d.).

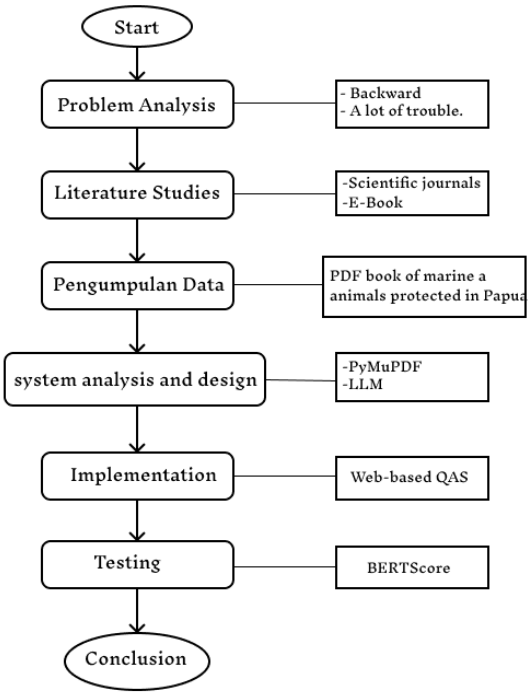


Figure 1. Research Stream

2.1 Problem Analysis

The first step in the research process is to identify the problem and understand its limits. Through this analysis, researchers are expected to understand the problems related to research in a comprehensive way in terms of functionality or help people in deepening their knowledge, thus gaining a better understanding. At this stage, the data used is a PDF book containing information about marine animals protected in Papua. The researchers corrected the data by converting the file to text format (txt) first, then re-copying the information to ensure the accuracy of the data (Dewi, Aunurohim, and Saptarini 2023).

2.2 Literature Studies

At this stage, researchers gather all the information needed for research. Researchers use a variety of methods to gather data that is relevant to the needs of ongoing research, including scientific journals, books, e-books, and websites (Homepage et al. 2024).

2.3 Data Collection

The purpose of the data collection process is to provide information that supports researchers in developing research as well as obtaining relevant and useful information to understand the phenomenon being studied, i.e. related to animals protected in Papua. Data collection is done by searching for a PDF book about animals protected in Papua, which is the validator of such data (Chatbot et al. 2022).

2.4 System Analysis & Design

This phase, which consists of a phase of investigation and identification of critical needs, is intended to meet application requirements. In this study, system requirements using PyMuPDF and Large Language paradigm (LLM) are used in system needs analysis to create a Question Answering System (QAS) that can offer answers to questions from PDF files and evaluate the application of such answers (Prasetyo, Benarkah, and Chrisintha 2021).

PyMuPDF is a Python module used to convert PDF files to text formats, and doc2text is used to extract text from other file formats. For cleaning, tokenization, stop-word removal, stemming, and lemmatisation. These steps help to do the text, reduce interference, and prepare it for further analysis (Eldi and Syaputra 2020).

A LLM is a newly created intelligence algorithm that is trained to anticipate the sequence of certain words based on the context of previous words. Generative Pre-trained Transformer (ChatGPT) belongs to the LLM category created by OpenAI and released in November 2022. In this study, the LLM plays a role in giving the highest value weight of some of the answers given by the system, where the answer with the greatest weight will be used as the answer to the user's question (Andesa 2019).

2.5 System Architecture Design

System architecture design for QAS Marine Animals protected in Papua using PyMuPDF and LLM. QAS system architecture is shown in the picture.

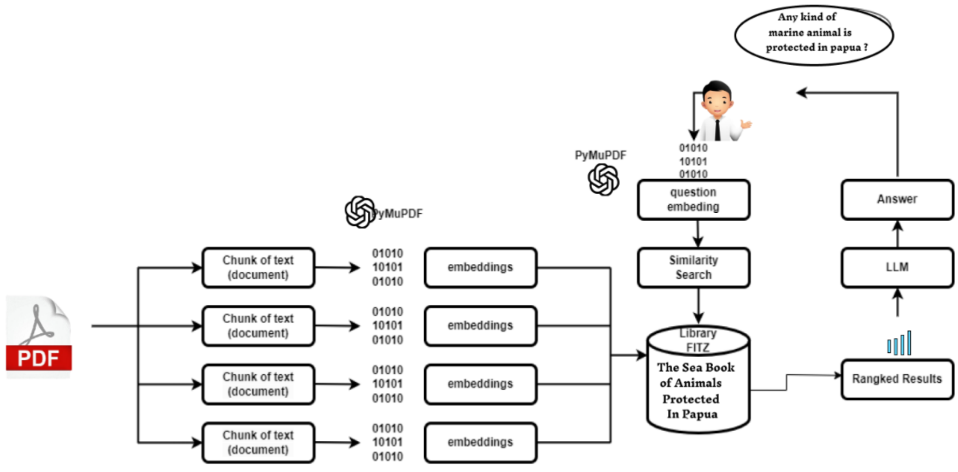


Figure 2. System Architecture Design

The initial stage in the development of this system is through the phase of preprocessing data, i.e. chunk of text and embeddings as in Figure 2 above. The phase is the process of grouping information into smaller parts to make it easier to understand and present information. In dividing sentences into larger sections, each section corresponds to a syntax unit such as an object phrase or a verb phrase. In the chunking process, the method only extracts keywords with the longest phrase. After that, all found keywords will be converted into keyword without any further selection or evaluation process. In other words, there is no further review or evaluation of the keywords that have been extracted. The size of the text cut is 1000 characters of the cut. Then the embedding is done, which is the process of converting words from human language into a machine-understood format, the 0101 numerical vector shape and stored in the PyMuPDF memory. In the embedding process, similar words will have representations of approximate or equal values. After the embedding process is completed, all texts (vectors) are stored in vector stores so that the Large Language Model (LLM) can recognize interlinkages. In this study, the researchers used a library called FITZ to store the vector stores. So for the data from the pdf book has become ready-to-use data that has become a knowledge base used as a database of the question answering system. If a user asks a question about the system, the question is answered using the embedding technique, which is then converted to a vector representation. After that, a common search using the PyMuPDF library was carried out in the vector stores to find some answers that match the question. After a few satisfactory answers, the answers are assessed according to their relevance, and then the LLM evaluates the answer to ensure that the most comprehensive answer is obtained (Rahayu, Harahap, and Agustian 2024).

2.6 Implementation

System implementation is steps to implement and run a designed system into everyday activities. It involves software installation, configuration, testing, and ensuring that the system works properly as needed (Sindy Nova, Nurul Khotimah, and Maria Y Aryati Wahyuningrum 2024).

1. **Web Development**

   In the development of the created system, the researchers used a streamlit framework. This framework was the choice because of its ability to facilitate the rapid and efficient development of interactive web applications using the Python programming language, without requiring a high level of expertise in the creation of complex user interfaces. By using streamlit, developers can focus on the implementation of the core features of the system being studied, while still providing a responsive and accessible user experience.

2.7 Testing

The purpose of this test is to assess whether the system that has been created can be used effectively or not. In this study, the test is carried out by testing whether the answers produced by the system correspond to the reference data already existing, i.e. using the method of Bidirectional Encodern Representations from Transformers (BERT) (Afriani et al. 2024).

1. ***BERTScore***

   BERTScore is a deep learning model that has provided significant progress in various natural language processing tasks. (NLP). BERT is designed to understand ambiguous sentences using the context of the surrounding text, building a more comprehensive understanding through the transformer. Transformer, with self-attention mechanisms, learns and adapts understanding to contextual relationships between words. BERT uses encoders in transformers as components in pre-training models for various natural language processing tasks such as Sentiment Analysis (SA), Question Answering (QA), and Text Summarization. (TS). Practically, BERT goes through two stages in its process, namely pre-training to understand language and fine-tuning for specific tasks. Here's the equation to calculate BERTScore.

   a. Precision

   Precision is the measure of how far the model prediction matches the requested data.

$$P_{BERT} = \frac{1}{|x|} \sum x_j \in x \max x \frac{T}{i} \hat{x} j \qquad (1)$$

   b. Recall

   Recall measures how effective models are in correctly predicting positive classes.

$$R_{BERT} = \frac{1}{x} \sum x_j \in x \max x \frac{T}{i} \hat{x} j \qquad (2)$$

   c. F-1 Score

   F-1 score F-1 score is the average value that compares precision and recall.

$$F_{BERT} = 2 \frac{PBERT\ .RBERT}{PBERT+RBERT} \qquad (3)$$

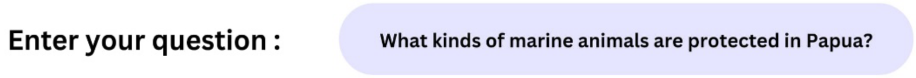
3. RESULT AND DISCUSSION

3.1 Implementation

The LLM based Chatbot is used to educate the knowledge of protected marine animals in Papua using GenerativeAI technology. Chatbots can interact in a personalised and responsive way to optimize knowledge about protected sea animals, thus effectively addressing a variety of user questions and issues. (Haryanto and Saefurrahman 2024).

1) Initial View

Users can start interacting with chatbots by asking about any kind of marine animal that is protected in Papua. Chatbots will respond by providing relevant content, using natural language processing to present information in an easy to understand way.



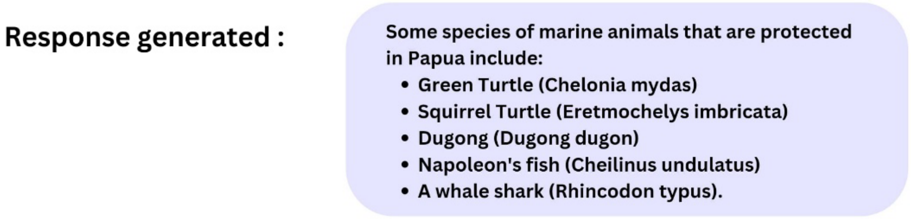
Enter your question :

What kinds of marine animals are protected in Papua?

Figure 3. Initial View

The image above is an initial interaction with a chatbot that shows how users start interacting with the chatbot to find out about protected marine animals in Papua.

2) Answer View
In this view, chatbots can provide structured and contextualized answers about protected marine animals in Papua. Chatbots combine information from data sets of marine knowledge that have been studied and use artificial intelligence to provide appropriate responses.



Response generated :

Some species of marine animals that are protected in Papua include:

- Green Turtle (Chelonia mydas)
- Squirrel Turtle (Eretmochelys imbricata)
- Dugong (Dugong dugon)
- Napoleon's fish (Cheilinus undulatus)
- A whale shark (Rhincodon typus).

Figure 4. Answer View

The above image shows the answer to a user's question showing how chatbots provide information about protected marine animals in Papua.

3.2 Testing

In this study, chatbot performance evaluation tests were conducted through a series of tests that compared chatbot responses to references of protected marine animals in Papua.

3.2.1 Chatbot Performance Evaluation

a. Evaluation Methods
The test was carried out by giving a series of questions to the chatbots covering different types of fish protected in Papua. Chatbots' responses were then evaluated based on the accuracy, clarity, and depth of the information provided.

Table 1. Chatbot Semantic Match Test Results

No	Question	Semantic Matching (%)
1.	Any kind of marine animal protected in Papua?	66%
2.	Why is the green turtle (Chelonia mydas) protected in Papua?	71%

3.	What is the role of dugong in the marine ecosystem in Papua?	69%
4.	How's the protection effort against the Napoleon fish (Cheilinus undulatus) in Papua?	69%
5.	What makes the whale shark (Rhincodon typus) a protected species in Papua?	75%
6.	What is the biggest threat to the survival of the Eretmochelys imbricata in Papua?	66%
7.	How do the local communities of Papua play a role in the conservation of marine animals?	78%
8.	What are the benefits of the conservation of protected marine animals in Papua?	89%
9.	How is the government's policy in protecting marine animals in Papua?	77%
10.	What can tourists do to support the conservation of marine animals in Papua?	69%

b. Result Analysis

Test results show that chatbots can consistently present accurate and relevant information about protected marine animals in Papua. A high degree of semantic compatibility indicates the ability of chatbot to process and answer questions accurately based on the data they have studied.

3.2.2 User Evaluation

User experience in using a chatbot is evaluated through satisfaction surveys to measure user responses and understanding of the chatbot.

a. User Satisfaction Survey

The results of the survey show that users feel that the chatbot is making a positive contribution to their understanding of protected marine animals in Papua.

Table 2. User Satisfaction Survey Results

No	Question	Satisfaction Presentation (%)
1.	I feel that the chatbot gives an informative answer.	86%
2.	Chatbots helped me understand the protected marine animals in Papua.	80%
3.	I'll use this chatbot again to find information about protected marine animals in Papua.	75%
4.	Chatbots are easy to use and responsive to my questions.	75%
5.	I feel like the chatbot has a deep enough knowledge about the protected marine animals in Papua.	75%

3.2.3 Evaluation From Testers

Feedback from this study highlights the potential of chatbots in providing accurate and relevant information about protected marine animals in Papua. As for what happens, the researchers suggest that chatbits are trained to analyze more complex questions over time and provide a more detailed context when discussing protected sea animals in Papúa.

3.3 Discourse

The research focuses on the development of web-based applications for interactive questions and answers sessions on protected marine animals in Papua. The development process includes data collection, text analysis, and evaluation of system performance with BERTScore metrics. The application was developed using AI, in particular the Big Language Model (LLM) and PyMuPDF algorithm, to provide accurate and relevant answers to user questions about protected sea animals populations in Papua. Initially, data was taken from related books in PDF format. Further, text editing, embedding, and vector storage creation processes were carried out to improve data input and retrieval. The implementation results are presented through a web search engine which makes it easier for users to find and check information about protected marine animals in Papua. The app's work quality assessment uses BERTScore metrics to determine the degree of semantic correspondence between the system and the reference response. The results show that the application can understand and answer questions about marine animals protected in Papua with a high degree of accuracy. The aim of this study is to provide users with easy, fast, and reliable access to information about protected marine animals in Papua.

4. CONCLUSION

The research focuses on developing educational chatbots with the aim of raising public awareness and understanding of protected marine animals in Papua. Using technologies such as Large Language Models (LLM) and PyMuPDF, the chatbot is designed to provide accurate and relevant information about rare marine species that are difficult to find. The development process includes problem analysis, literature study, data collection, system analysis and system design.

The results of this study show that the system can provide users with relevant and comprehensive information as well as identifying the types of marine animals protected in Papua. Nevertheless, the study also identifies some areas that still need to be improved in the coming years. The first area in question is improved work productivity and system functionality, through the integration of new technologies such as more sophisticated manufacturing processes or more thorough data analysis. Further research is recommended to improve the quality and consistency of data in case studies so that it is more relevant to user questions.

REFERENCES

Aditya, Zaka Firma, and Sholahuddin Al-Fatih. 2017. “Perlindungan Hukum Terhadap Ikan Hiu Dan Ikan Pari Untuk Menjaga Keseimbangan Ekosistem Laut Indonesia.” *Jurnal Ilmiah Hukum LEGALITY* 24 (2): 224. https://doi.org/10.22219/jihl.v24i2.4273.

Afriani, Elvina, Nazruddin Safaat H, Muhammad Fikry, and Muhammad Affandes. 2024. “APLIKASI TANYA JAWAB TENTANG FIQIH BERSUCI BERBASIS WEB” 6 (2): 380–90.

Andesa, Khusaeri. 2019. “Super Agent Chatbot ‘3S’ Sebagai Media Informasi Menggunakan Metoda Natural Language Processing (NLP).” *Jurnal Teknologi Dan Open Source* 2 (1): 53–64.

Bawole, Roni, and Rony Megawanto. 2017. “Establishing of Aquatic Protected Areas (Apas) Network in Papua’S Bird Head’S Seascape (Bhs): Species Migration and Genetic Connectivity.” *Coastal and Ocean Journal (COJ)* 1 (2): 189–200. https://doi.org/10.29244/coj.1.2.189-200.

Chatbot, Pengembangan, Pengaduan Dan, Teknologi Informasi, Dengan Pendekatan, Studi Kasus, and Politeknik Negeri. 2022. “1028-Article Text-4957-1-10-20221228” 12 (2): 575–83.

Dewi, Shinta Cahyaning, Aunurohim, and Dian Saptarini. 2023. “Karakteristik Mikroplastik Pada Ikan Kakatua Anglu (Chlorurus Sordidus) Dan Ikan Kurisi Sirip Pucat (Nemipterus Thosaporni) Di Perairan Teluk Jakarta.” *Jurnal Kelautan* 16 (3): 2476–9991.

Eldi, Eldi, and Hadi Syaputra. 2020. “Implementasi Chatbot Untuk Mendukung Sistem Informasi Pada Rumah Sakit Muhamadiyah Palembang.” *Jurnal Nasional Ilmu Komputer* 1 (3): 139–48. https://doi.org/10.47747/jurnalnik.v1i3.160.

Haryanto, Iqbal Dwi, and Saefurrahman Saefurrahman. 2024. “Implementasi Chatbot Kesehatan Kucing Melalui Dialogflow Dan Telegram Untuk Pemberian Informasi Penyakit Dan Perawatan.” *JTIM : Jurnal Teknologi Informasi Dan Multimedia* 5 (4): 365–76. https://doi.org/10.35746/jtim.v5i4.484.

Homepage, Journal, Anggun Tri Utami Br Lubis, Nazruddin Safaat Harahap, Surya Agustian, Muhammad Irsyad, Iis

Afrianty, Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, and Corresponding Author. 2024. "MALCOM: Indonesian Journal of Machine Learning and Computer Science Question Answering System on Telegram Chatbot Using Large Language Models (LLM) and Langchain (Case Study: Health Law) Question Answering System Pada Chatbot Telegram Menggunakan Large Language Models (LLM) Dan Langchain (Studi Kasus UU Kesehatan)" 4 (3): 955–64.

Informatika, Prodi Teknik, and Universitas Abdurrab. n.d. "O c p n l P," 17–26.

Prasetya, Sigit Heru, Ambariyanto Jurusan, Ilmu Kelautan, and Fakultas Perikanan. 2014. "Estimasi Daya Dukung Terumbu Karang Berdasarkan Biomassa Ikan Karang Di Perairan Misool Selatan, Raja Ampat, Papua Barat." *Journal of Marine Research* 3 (3): 233–43.

Prasetyo, Vincentius Riandaru, Njoto Benarkah, and Vioni Jannet Chrisintha. 2021. "Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya." *Teknika* 10 (2): 114–21. https://doi.org/10.34148/teknika.v10i2.370.

Rahayu, Suci, Nazruddin Safaat Harahap, and Surya Agustian. 2024. "Application of Langchain Technology to the Fiqh Question Answering System of Four Madhhab Penerapan Teknologi LangChain Pada Question Answering System Fikih Empat Madzhab" 4 (July): 974–83.

Ramadhan, Helmi, Shifa Helena, and Yusuf Arief Nurrahman. 2024. "Struktur Komunitas Ikan Kakatua (Scaridae) Di Selatan Pulau Kabung Kabupaten Bengkayang, Kalimantan Barat." *Jurnal Laut Khatulistiwa* 7 (1): 1. https://doi.org/10.26418/lkuntan.v7i1.63255.

Sindy Nova, Nurul Khotimah, and Maria Y Aryati Wahyuningrum. 2024. "Pemanfaatan Chatbot Menggunakan Natural Language Processing Untuk Pembelajaran Dasar-Dasar Gui Tkinter Pada Bahasa Pemrograman Python." *Jurnal Ilmiah Teknik* 3 (1): 58–65. https://doi.org/10.56127/juit.v3i1.1162.