JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

Lightweight Retrieval-Augmented Generation Based on Google Sheets as a Dynamic Knowledge Base for the Maritime BMKG Teluk Bayur Shipping Safety Chatbot

Andini Nareswari Rifaat[1*], Budi Iman Samiaji[2]

[1]Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada, Indonesia
[2]Teluk Bayur Maritime Meteorology Station, Badan Meteorologi Klimatologi dan Geofisika, Indonesia
[1*]andininareswaririfaat@mail.ugm.ac.id
*Corresponding Author

***Abstract*--Maritime weather information is critical for shipping safety in archipelagic countries such as Indonesia. However, current dissemination methods at the Teluk Bayur Maritime Meteorological Station rely on one-way infographic distribution through social media, lacking interactive communication capabilities. This study presents a lightweight Retrieval-Augmented Generation (RAG)-based chatbot that integrates Google Sheets as a dynamic knowledge base for real-time maritime weather dissemination through a conversational interface. The system was implemented using Python and Streamlit within a four-layer monolithic architecture and covers six information categories, including early warnings, three-day weather summaries, and vessel safety recommendations across nine maritime regions in West Sumatra for four vessel types. A rule-based decision support module classifies sailing conditions as safe *(aman)* or cautionary *(waspada)* based on wind speed and wave height thresholds defined by Meteorological, Climatological and Geophysics Agency of Indonesian (BMKG). Functional testing confirmed accurate responses across all menu features and free-form queries. Quantitative evaluation using 50 queries, scored on a 0–1 scale using ROUGE-L and TF-IDF cosine similarity, showed that the RAG-enabled model achieved an overall mean score of 0.2285, compared to 0.1425 for the non-RAG baseline (+0.0860) demonstrating that structured context grounding effectively reduces hallucination and enhances information reliability for maritime safety applications.**

***Key words*: Chatbot; Google Sheets; Maritime weather; Retrieval-augmented generation; Safety shipping.**

I. INTRODUCTION

Maritime weather information is a critical public safety service in archipelagic countries such as Indonesia. The Class IV Teluk Bayur Maritime Meteorological Station, operated by the Meteorological, Climatological, and Geophysical Agency of Indonesia (BMKG), monitors nine maritime areas in West Sumatra and issues daily wave-height forecasts, wind advisories, and early-warning bulletins. To ensure that this vital information reaches maritime users effectively, BMKG disseminates these updates through an integrated digital network [1], including its official web portal, Instagram, and dedicated WhatsApp groups.

Maritime weather information is essential for seafarers to prevent maritime accidents. The severity of this risk is illustrated by recent maritime accident in West Sumatra and the Mentawai Islands. On March 5, 2025, a vessel lost contact after being struck by a severe storm near Pitojat Island [2]. This was followed by another tragic incident on July 14, 2025, when a longboat carrying 18 passengers capsized in the Sipora Strait due to severe weather conditions [3]. These incidents underscore the critical importance of BMKG's early warning services.

In many cases, these tragedies are not merely the result of nature's volatility, but also of a critical lack of understanding of maritime conditions. This is supported by research analyzing a decade of global accident reports, which identifies a significant correlation between environmental severity, such as strong winds and high waves, and human factors [4]. Specifically, the study highlights that seafarers with limited sailling experience or insufficient theoretical knowledge are more vulnerable to serious incidents [4]. These findings suggest that when adverse weather conditions coincide with limited professional preparedness or inadequate weather literacy, the risk of a high-severity maritime accident increases substantially [4].

The BMKG employs various scientific tools and methods to support the production and dissemination of weather and climate forecast information in Indonesia [1]. Through AI-driven systems built on more than 30 years of historical meteorological data, including temperature,

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

humidity, and atmospheric parameters, BMKG is able to deliver high-precision hazard predictions that support critical national sectors [5]. For example, in public health, these systems enable the early prediction of dengue fever (DBD) risk zones up to a week in advance [5]. In agriculture, the spatial predictions with a resolution of up to 3 km helps farmers determine optimal planting periods [5]. Meanwhile, in logistics and transportation, AI facilitates real-time monitoring of weather conditions along distribution routes, thereby improving the safety and efficiency of essential goods delivery [5].

Various studies have been conducted to improve the dissemination of weather information, particularly maritime weather information. To modernize these services, Retrieval-Augmented Generation (RAG) has emerged as a transformative bridge between complex meteorological data and the practical needs of seafarers. One notable example is the MeteoChat framework, which utilizes a RAG pipeline to combined with LangChain, fine-tuning, and chatbot implementation semi-automate the generation of environmental reports, thereby improving the speed and clarity of maritime weather information dissemination [6].

Recent developments in chatbot architecture indicate a paradigm shift from traditional linear pipelines toward more flexible and scalable frameworks. The “Modular RAG” approach introduces a systematic decomposition of chatbot systems into reconfigurable modules, in which independent components and specialized operators enable advanced mechanisms such as dynamic routing, scheduling, and information fusion, thereby extending beyond the conventional retrieve-then-generate paradigm [7]. This transformation redefines chatbot systems as composable architectures capable of supporting branching, conditional, and iterative execution flows. Complementing this conceptual advancement, microservice-based implementations further address the scalability and maintainability limitations of monolithic designs by providing reusable and production-oriented architectural blueprints [8]. Collectively, these approaches demonstrate the ongoing evolution of RAG-based chatbots from proof-of-concept systems into robust, scalable, and enterprise-ready solutions.

Despite the growing emphasis on modular and microservice-based architectures, many practical implementations still adopt cloud-hosted monolithic designs because of their simplicity and ease of deployment [9]. This trend is particularly relevant for lightweight applications that integrate external data sources and large language models through APIs.

In line with the practical advantages of cloud-hosted monolithic deployments, this study presents a maritime weather chatbot implemented as a lightweight RAG system that integrates external APIs within a unified application developed using Streamlit. The main contributions of this study are as follows: (1) the design of a lightweight RAG architecture that utilizes live Google Sheets data as a dynamic knowledge base; (2) the development of a rule-based decision support module derived from BMKG maritime safety advisories to generate sailing condition classifications (safe and caution) for various vessel types; (3) the implementation of a functional chatbot framework to facilitate the dissemination of maritime weather information dissemination in West Sumatra at BMKG Teluk Bayur; and (4) an empirical evaluation using 50 tes queries, demonstrating high response accuracy.

II. Method

A. Requirement Analysis

The initial phase of this research involved identifying the requirements for a chatbot system designed to disseminate maritime weather information at the Teluk Bayur Maritime Meteorological Station. An analysis was conducted of the characteristics of the data used, namely spreadsheet-based maritime weather data that are updated at regular intervals, as well as the needs of users, such as fishermen and vessel operators, who require information that is timely, accurate and easy to understand.

In addition, the limitations of the existing system were identified, including the presentation of weather information in the form of infographics disseminated through social media and the lack of two-way interaction. Based on this analysis, the system requirements were formulated as a Retrieval-Augmented Generation (RAG)-based chatbot capable of integrating real-time data, generating natural responses, and providing simple decision support regarding maritime safety with six main categories of user information needs, namely:

1. Maritime weather early warning information.
2. Three-day weather forecast summaries by maritime region.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

3. Maritime safety recommendations based on vessel type.
4. Sea surface temperature information by maritime region.
5. Historical weather data queries based on specific dates
6. Information on BMKG location, contact details, and official weather infographics.

B. System Architecture

The proposed system employs a lightweight RAG architecture, in which all system components are integrated into a single application. This architecture differs from conventional RAG approaches that rely on vector databases such as FAISS or Pinecone [10]. Instead, the proposed system utilizes Google Sheets as a dynamic knowledge base that can be updated directly by BMKG staff on a daily basis.

The architecture consists of four layers, as shown in Fig. 1. **First**, the presentation layer: a Streamlit-based chat interface with a fixed header, containing the main menu, user text input, and reply bubbles from the LLM. **Second**, the data layer, which utilizes Google Sheets through the Google Visualization API as a dynamic knowledge base storing maritime weather information. Data is cached for 600 seconds using the *@st.cache*_data decorator to minimize API calls whilst maintaining the freshness of operational data. **Third**, the processing layer, which covers data retrieval, preprocessing, and context formation for the language model. Data from Google Sheets is retrieved directly, then converted into a structured format (JSON/tabular) to be used as context in the response generation process. **Fourth**, the generation layer, uses a Large Language Model (LLM) via the OpenAI API to generate responses based on the retrieved contextual information. In addition, a rule-based decision support module is incorporated to classify sailing conditions as either safe or cautionary based on predefined weather parameters.

The chatbot is designed using a menu-based conversational flow approach [11], in which users can choose from six main menus or type in a free-form question. The menu-based approach was chosen because it minimizes intent ambiguity while providing clear navigation guidance for novice users. The list of user intents and the corresponding system responses for each menu is presented in Table I.

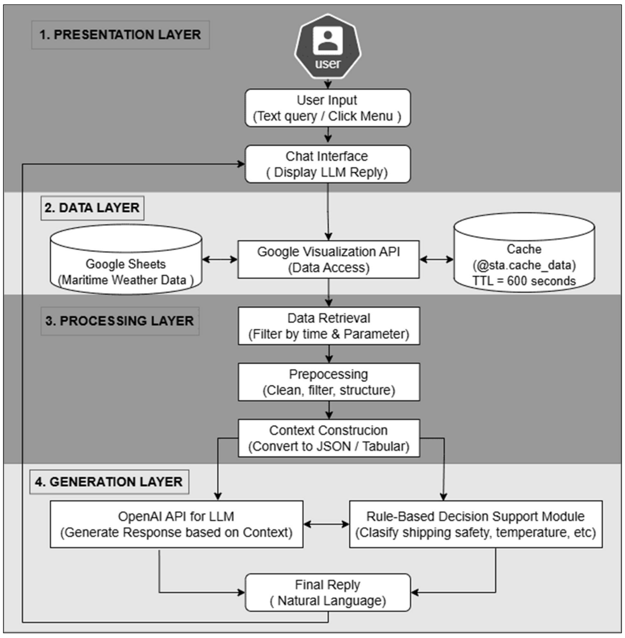

Fig. 1. System Architecture

TABLE I
List of User Intents and Responses

Intent / Menu & User Input	System Response
Early Warning Info *Menu click or free-text query*	3-day early warning for West Sumatra sea areas, filtered from Sheet "*Narasi*" (rows 35 - 49), excluding tidal extremes
3-Day Weather Summary *Menu click or free-text query*	Per-area summary: significant weather, wind direction/speed, wave height, sea temperature, humidity, and current sourced from Sheet "*Ringkasan Semua*"
Vessel Safety Recommendation *Menu click or free-text query*	Color-coded HTML table showing SAFE / WARNING status for 4 vessel types across 9 sea areas, generated by the rule-based safety module
Sea Temperature Info *Menu click or free-text query*	Color-coded sea surface temperature (°C) per area for today, tomorrow, and the day after sourced from Sheet "*Ringkasan Semua*"
Infographics & Website *Menu click*	Links to BMKG official website, maritime portal, and Instagram @stamar_tlkbayur
Location & Contact *Menu click*	Station address, WhatsApp link, Instagram, and Google Maps link
Free-text weather query *Natural language (e.g., "How are the waves in Sipora today?")*	LLM-generated response grounded in retrieved tabular data for the detected intent and date
Historical date query *Natural language with date (e.g., "What was the weather on 15/03/2026?")*	LLM-generated response using data fetched for the specified date from Sheet "*Ringkasan Semua*"

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

Intent / Menu & User Input	System Response
Out-of-scope query *Any non-maritime topic*	Polite refusal; system redirects user to maritime weather topics and provides contact information
Unrecognized input *Unmatched or ambiguous text*	LLM display today weather condition

C. Implementation

The system was implemented using the Python programming language, with the Streamlit framework serving as the user interface. The system is integrated with Google Sheets through an API to retrieve data in real time. The main processes of the system are as follows:

- Data Retrieval
  Maritime weather data are retrieved from Google Sheets and filtered based on the date and parameters specified by the user.
- Intent Detection and Query Parsing
  The system identifies the user's intent based on keywords such as temperature, wind, or waves, and extracts temporal information, including today, tomorrow, or a specific date.
- Context Construction (RAG)
  The filtered data are transformed into a structured textual representation that is incorporated into the language model prompt.
- Response Generation
  The language model generates a response based on the retrieved contextual data and predefined system rules.
- Decision Support Module for shipping safety
  The system applies rules-based criteria derived from wind speed and wave height thresholds to classify sailing conditions as either safe or cautionary for different types of vessels.

D. Evaluation

The evaluation was conducted through two approaches: functional testing and comparative performance analysis. Functional testing was performed via direct interaction with the chatbot user interface to verify system functionality across all features, while natural language queries were used to simulate realistic user behavior.

For quantitative evaluation, a total of 50 queries covering maritime weather information and maritime safety scenarios were evaluated using two configurations: the proposed Retrieval-Augmented Generation (RAG) model and a baseline non-RAG model. Ground-truth responses were constructed from official maritime weather data provided by BMKG through Google Sheets and validated by a professional weather forecaster. The evaluation was performed automatically using an NLP evaluation algorithm; among these, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics family (i.e., ROUGE-L, ROUGE-W, and ROUGE-S) [12] proved to be very effective in the automatic evaluation of machine translation [13] with rouge-score python packages. In addition, the Term Frequency-Inverse Document Frequency (TF-IDF) metric, which is a fundamental tool for understanding the relevance of words in documents [14] with scikit-lear python packages.

As shown in Table II, there are four criteria were used: accuracy (factual correctness), relevance (alignment with the user query), clarity (readability), and completeness (information coverage). Each criterion was scored on a scale ranging from 0 to 1. The overall score was calculated as the average of these four metrics. This evaluation framework follows recent practices in LLM-based assessment [15], enabling an objective comparison between RAG and non-RAG approaches.

TABLE II
NLP Metrics Evaluation

Criteria	Metrics	Library
Accuracy	ROUGE-L Precision	rouge-score
Relevance	TF-IDF Cosine Similarity	scikit-learn
Clarity	ROUGE-L F1	rouge-score
Completeness	ROUGE-L Recall	rouge-score

Answer quality was evaluated using two complementary metrics. Accuracy, clarity, and completeness were measured using ROUGE-L [13], whereas relevance was evaluated using TF-IDF-weighted cosine similarity [14], [16]. ROUGE-L is based on the Longest Common Subsequence (LCS) between a reference sentence $X$of length $m$and a candidate sentence $Y$of length n[13]. The LCS-based recall, precision, and F-measure are defined as [13]:

$$R_{lcs} = \frac{LCS(X,Y)}{m}, P_{lcs} = \frac{LCS(X,Y)}{n}, F_{lcs} = \frac{(1+\beta^2)\, R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad (1)$$

where $LCS(X,Y)$is the length of the longest common subsequence of $X$and Y, and $\beta = 1$, giving equal weight to precision and recall. P_{lcs}, F_{lcs}, and R_{lcs}were

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

used in this study as proxies for accuracy, clarity, and completeness, respectively.

TF-IDF is to measure relevance, each text was first represented as a Term Frequency–Inverse Document Frequency (TF-IDF) weighted vector. The weight of term $t$in document $d$within corpus $D$is given by [14]:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \times \log_{10} \frac{|\,D\,|}{|\,\{d \in D : t \in d\}\,|} \quad (2)$$

Given the TF-IDF vector of the reference text $\vec{r}$ and the candidate text $\vec{c}$, each with N dimensions corresponding to the vocabulary size of the corpus, relevance was computed as the cosine similarity between the two vectors [16]:

$$\text{cosine}(\vec{r},\vec{c}) = \frac{\vec{r}\cdot\vec{c}}{\|\,\vec{r}\,\|\ \ \|\,\vec{c}\,\|} = \frac{\sum_{i=1}^{N} r_i\, c_i}{\sqrt{\sum_{i=1}^{N} r_i^2}\ \sqrt{\sum_{i=1}^{N} c_i^2}} \quad (3)$$

where $r_i$and $c_i$denote the TF-IDF weight of the i-th term in the reference and candidate vectors, respectively.

All four scores (P_{lcs}, F_{lcs}, R_{lcs}, cosine) range from 0 to 1, with higher values indicating greater similarity to the ground truth. ROUGE-L scores were computed using the rouge-score Python library, while TF-IDF vectorization and cosine similarity were computed using scikit-learn.

III. RESULT AND DISCUSSION

A. *Functional Testing Results*

Functional testing was conducted to evaluate the system's ability to respond accurately to predefined menu selections as well as free-form user queries. The chatbot was tested across all available menu options, and its responses were evaluated in terms of correctness, completeness, and relevance.

As present in the Fig. 2, the main interface provides six interactive menu buttons: *Informasi Peringatan Dini*, *Layanan Infografis dan Website*, *Ringkasan Cuaca Maritim Sumbar (3 Hari)*, *Jenis Perahu yang Diizinkan Berlayar*, *Lokasi dan Kontak*, and *Informasi Suhu Maritim*. Each button triggers a structured response retrieved from the system's knowledge base. This design reduces input ambiguity and enables users to access frequently requested maritime weather information without requiring prior knowledge of query syntax [17].

When the "**Informasi Peringatan Dini**" menu was activated likes figure 3, the system successfully returned a three-day early warning bulletin covering West Sumatera waters, dated April 23–25, 2026. The response included medium wave height warnings *(Peringatan Gel. Sedang)* of up to 2.5 meters across multiple maritime zones, along with rainfall potential alerts. This output is consistent with the structured bulletin format issued by BMKG[18]. The "**Ringkasan Cuaca Maritim**" Sumbar (3 Hari) menu returned a three-day summary encompassing significant weather conditions, wind speed and direction, wave heights, sea surface temperature, humidity, and current data demonstrating the system's capacity to present multi-parameter meteorological information in a structured conversational format as present in the Fig. 4.

The "**Informasi Suhu Maritim**" feature displayed a tabular sea surface temperature breakdown across nine sea zones for three consecutive dates likes in the Fig. 5 and this condition have values ranging between 27.4°C and 28.1°C. This structured table presentation within a chat interface follows patterns documented in weather information chatbot implementations that integrate database-driven response generation [19].

The "**Saran Keselamatan Perahu untuk Berlayar**" feature generated a maritime safety recommendation table (*Rekomendasi Keselamatan Pelayaran)*, classifying sailing conditions as either *WASPADA* (caution) or *AMAN* (safe) four classes of vessels across nine maritime zones over a three-day period, as shown in Fig. 6. The recommendations provide actionable guidance for mariners based on predefined thresholds of maximum wind speed and wave height [20].

As shown in Fig. 7, the "**Layanan Infografis dan Website**" menu returned links to the official BMKG Maritime website, station website, and Instagram account, while the "**Lokasi dan Kontak**" menu provided complete contact details including WhatsApp number, Google Maps reference, and physical address confirming accurate delivery of institutional information. Fig. 8 illustrates the system's multi-turn free-form query handling, where the chatbot correctly identified *WASPADA* conditions in Timur Siberut with wave heights of 1.8 meters and wind speeds of 5 knots, recommending voyage postponement, and subsequently confirmed *AMAN* status for ferry vessels across all maritime zones in a follow-up query demonstrating the system's ability to maintain conversational context and deliver data-driven responses.

As an additional evaluation, Fig. 9 present one of 50 queries comparing ground truth, LLM with RAG, and LLM Non-RAG responses. The ground-truth has been annotated by the forecaster

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

with a "✓" mark. The question asked relates to marine fog, where the response given by the LLM based on RAG is more accurate because it presents data related to weather conditions and directly states that fog-related information is not available. Meanwhile, the response from the LLM Non-RAG model is more biased and narrative without providing real-world conditions.

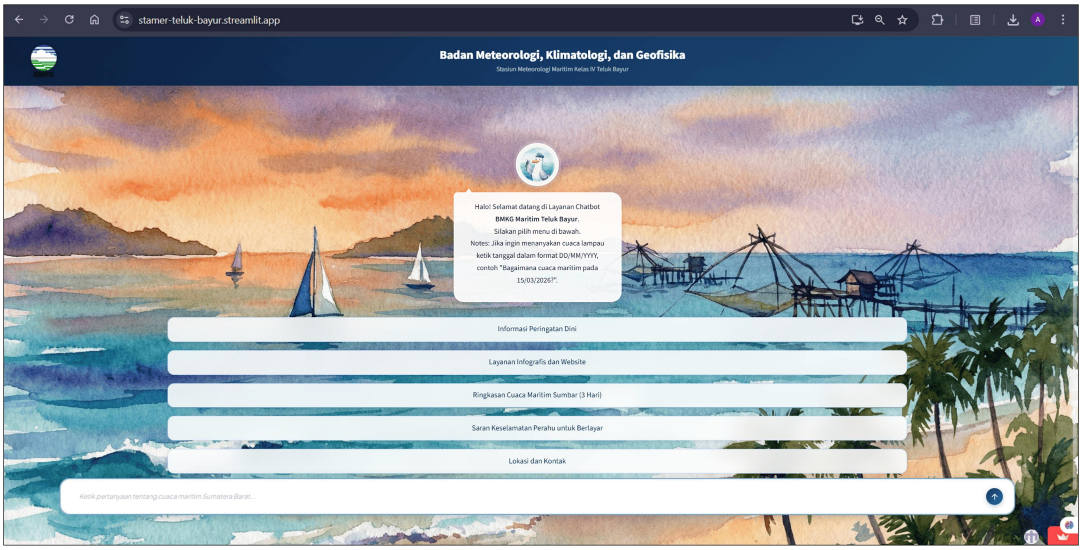

Fig. 2. Chatbot Menu

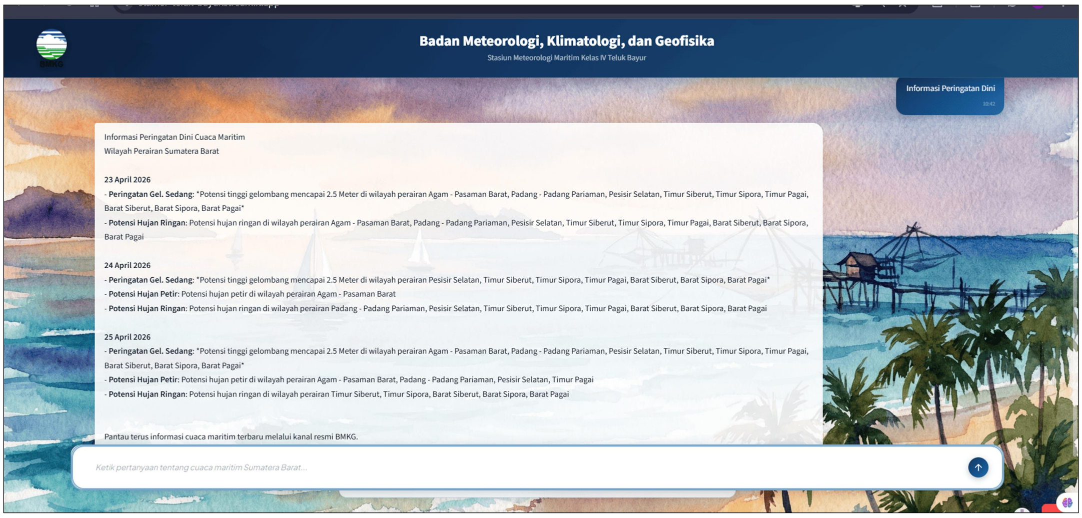

Fig. 3. Early Warning Menu

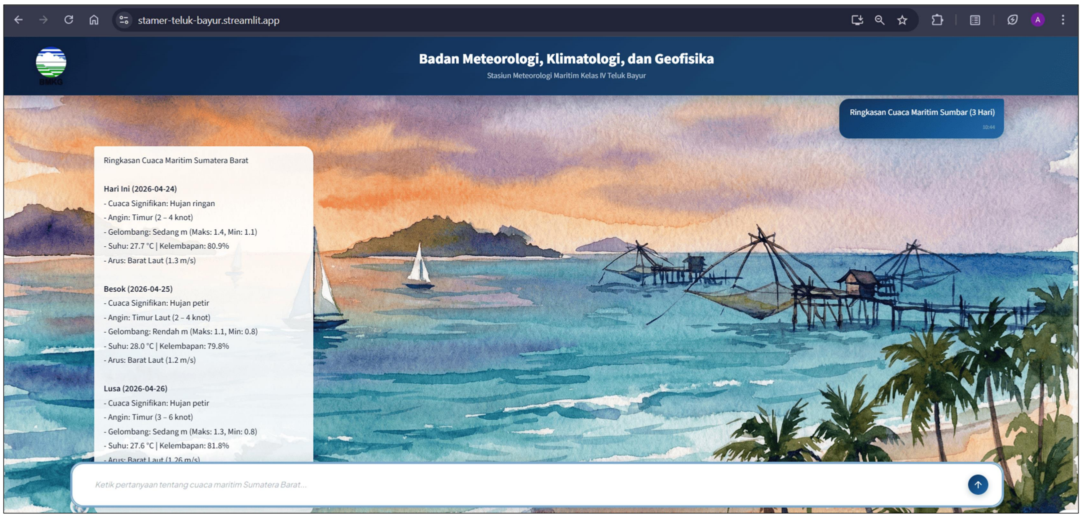

Fig. 4. Weather Summaries Menu

JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

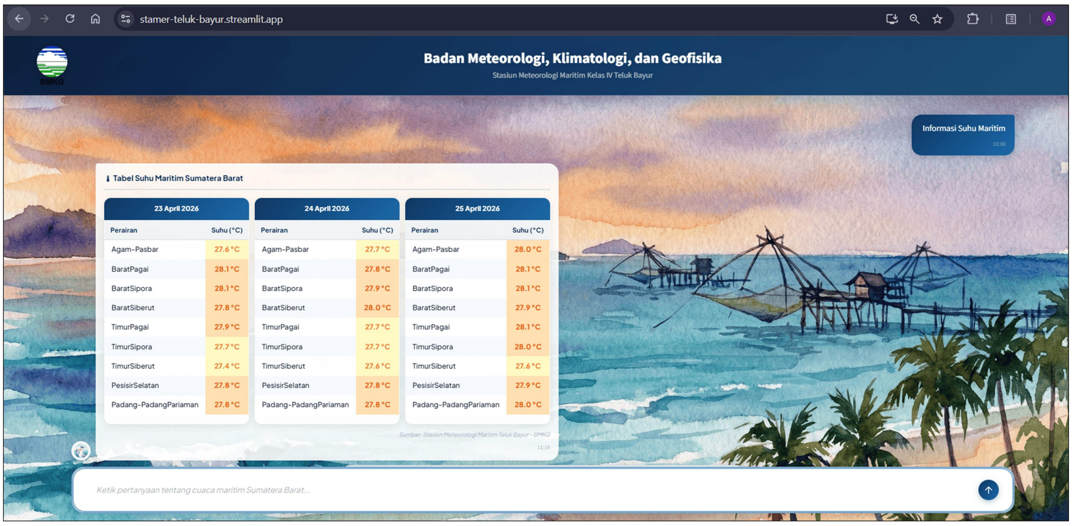

Fig. 5. Temperature Information Menu

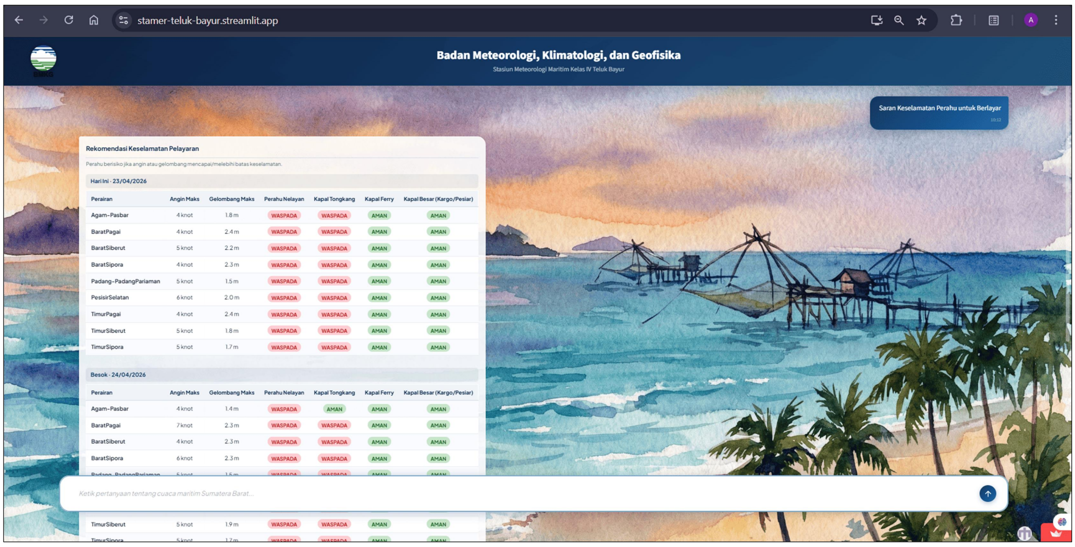

Fig. 6. Shipping Safety Menu

Fig. 7. Contact, Location, and Infographics Menu

JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

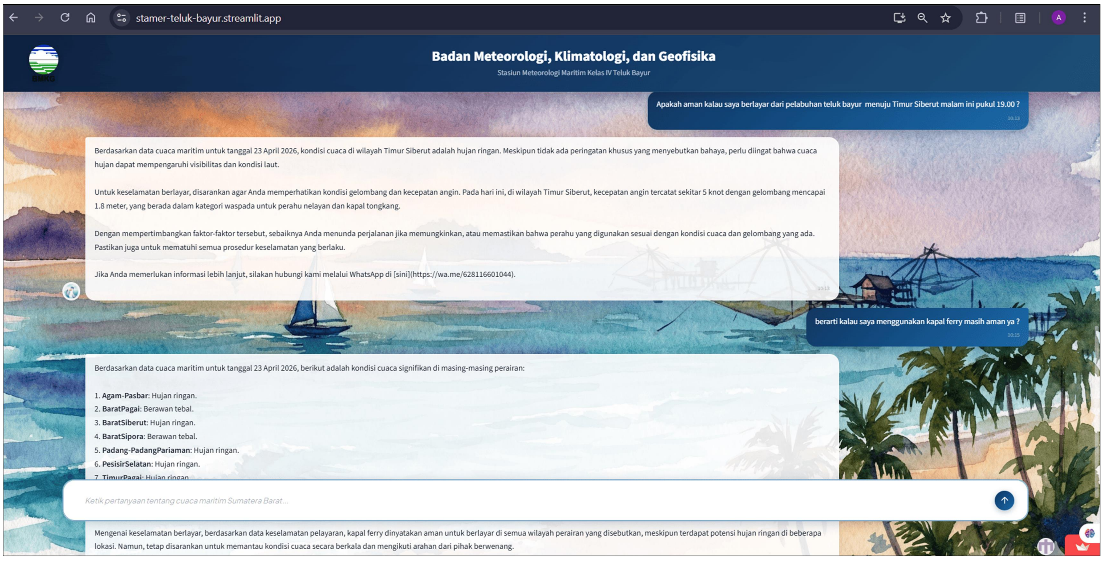

Fig. 8. Natural questions

no	query	ground_truth	rag_answer	non_rag_answer
6	Di Timur Pagai ada kemungkinan berkabut dalam waktu dekat nggak?	Berdasarkan data yang tersedia, untuk perairan Timur Pagai pada hari ini (21 April 2026), kondisi cuaca yang terpantau adalah "Hujan ringan" dengan suhu rata-rata 27.4°C dan kelembapan relatif (RH) 80.3%. Kecepatan angin maksimum tercatat 6 m/s dan tinggi gelombang maksimum mencapai 1.5 meter. Data mengenai kemungkinan berkabut tidak tersedia dalam informasi. ✓	Berdasarkan data cuaca maritim untuk tanggal 21 April 2026, di wilayah Timur Pagai terpantau cuaca dengan kondisi hujan ringan. Meskipun tidak ada informasi spesifik mengenai kemungkinan berkabut, hujan ringan dapat mempengaruhi visibilitas. Oleh karena itu, disarankan untuk tetap waspada dan memantau perkembangan cuaca lebih lanjut. Jika Anda memerlukan informasi lebih lanjut, silakan hubungi kami melalui WhatsApp di https://wa.me/628116601044 atau IG @stamar_tlkbayur.	Untuk wilayah Timur Pagai, kondisi cuaca dapat dipengaruhi oleh berbagai faktor, termasuk kelembapan dan pola angin. Secara umum, kabut dapat terjadi pada malam dan pagi hari, terutama jika suhu udara turun dan kelembapan tinggi. Namun, untuk informasi yang lebih akurat dan terkini mengenai kemungkinan kabut di Timur Pagai, disarankan untuk memantau laporan cuaca dari BMKG atau sumber resmi lainnya. Pastikan juga untuk selalu memperhatikan peringatan cuaca yang dikeluarkan, terutama jika Anda berencana untuk beraktivitas di laut. Jika ada pertanyaan lebih lanjut atau membutuhkan informasi lain, silakan sampaikan.

Fig. 9. Sample of responses comparison RAG vs. Non-RAG

B. *Quantitative Evaluation between RAG and Non-RAG Model*

A total of 50 test queries were evaluated across both system configurations using four metrics, each scored on a scale of 0–1. A summary of results is presented in Table III. Those four matrices consisted of ROUGE-L precision for accuracy, TF-IDF cosine similarity for relevance, ROUGE-L F1 for clarity, and ROUGE-L recall for completeness, comparing the answers generated by the RAG and non-RAG systems against the ground truth. The results indicate that the RAG based system achieved a higher mean overall score (M = 0.2285) compared to the non-RAG system (M = 0.1425), corresponding to a mean difference of 0.0860 on a 0-1 scale.

This advantage was consistent across all four metrics. For accuracy, RAG outperformed non-RAG with mean ROUGE-L precision scores of 0.2681 in comparison with 0.1390, respectively. A similar pattern emerged for relevance, where RAG achieved a higher mean TF-IDF cosine similarity (0.1768) than non-RAG (0.1030). In terms of clarity, RAG attained a mean ROUGE-L F1 score of 0.2268 compared to 0.1458 for non-RAG, while for completeness, RAG scored 0.2422 against 0.1824 for non-RAG on the ROUGE-L recall metric. Taken together, these results suggest that the RAG architecture consistently produced answers more closely aligned with the ground truth across all evaluated dimensions.

At the query level, RAG outperformed non-RAG on all 50 queries, with no instances of a tie or a non-RAG win. This consistent performance across the entire evaluation set lends further support to the conclusion that the retrieval component meaningfully contributes to answer quality, rather than the observed improvement being driven by a small subset of favorable cases.

Despite this consistent advantage, the absolute scores achieved by both systems remained relatively low (below 0.30 across all four metrics). This outcome is partly attributable to the nature of the evaluation metrics employed: ROUGE-L and TF-IDF cosine similarity are lexical overlap-based

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

measures that are sensitive to exact word matching and word order, and therefore tend to penalize semantically correct answers that are phrased differently from the ground truth. Since the chatbot was designed to generate natural, conversational responses rather than verbatim reproductions of the reference text, a certain degree of lexical divergence from the ground truth was expected. Consequently, the relative improvement of RAG over the non-RAG baseline is considered a more meaningful indicator of system performance in this study than the absolute magnitude of the scores.

TABLE III
Quantitative Comparison Results

Criteria	RAG Score	Non-RAG Score	Improvement
Accuracy	0.2681	0.1390	+0.1291
Relevance	0.1768	0.1030	+0.0738
Clarity	0.2268	0.1458	+0.0810
Completeness	0.2422	0.1824	+0.0598
Overall Score	0.2285	0.1425	+0.0860

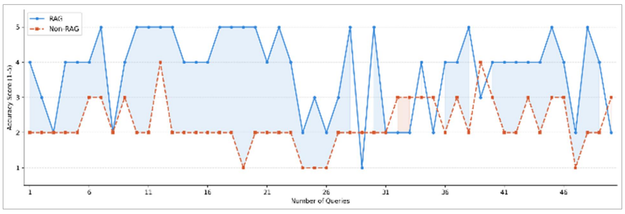

Fig. 10. Per-query accuracy scores for RAG vs Non-RAG

IV. Conclusion

This study developed a lightweight RAG-based chatbot for disseminate maritime weather and safety information from BMKG Teluk Bayur, complementing the existing one-way infographic dissemination system. The proposed system provides an interactive conversational interface supported by a dynamic Google Sheets-based knowledge base. It covers six categories of information with a rule-based safety classification module for four types of vessels across nine maritime regions. Functional testing confirmed that all system features operated correctly and were capable of responding to both predefined menu selections and free-form user queries. A quantitative evaluation of 50 queries showed that the RAG model outperformed the non-RAG baseline with an overall score of 4.04 vs. 3.11 (+0.93), yielding better responses for 78% of queries, with the greatest improvements in accuracy (+1.52) and relevance (+1.08), confirming that the answers generated by RAG are more accurate and have less hallucinations than the answers from Non-RAG.

V. Acknowledgment

The authors gratefully acknowledge the Stasiun Meteorologi Maritim Kelas IV Teluk Bayur, Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), for providing access to operational maritime weather data and for its institutional support throughout this research.

VI. References

[1] Raja. P. Siregar, M. Fadli, and M. Khaerunnisa, Modul Pelatihan Informasi Cuaca Maritim untuk Nelayan Tangkap Adaptasi Perubahan Iklim dan Ketangguhan (APIK). Stasiun Meteorologi Juanda, 2017. Accessed: Apr. 21, 2026. [Online]. Available: https://stamet-juanda.bmkg.go.id/home/data/buletin/Modul-PICM-Nelayan-Tangkap.pdf?__cf_chl_tk=_Oz1gdsC40eF1E5n917uk_i8ijZehhmzLuOdw1ThiK4-1776687813-1.0.1.1-2U.aOp.szJUb4XhtIyDHElHePQP5PmFKM4GfmTMas7I#page=4.08

[2] "Perahu Nelayan Hilang Kontak setelah di Hantam Badai di Perairan Pulau Pitojat - Website Resmi Dinas Kominfo Kabupaten Kepulauan Mentawai." Accessed: Apr. 21, 2026. [Online]. Available: https://diskominfo.mentawaikab.go.id/2025/03/perahu-nelayan-hilang-kontak-setelah-di.html

[3] Y. Sastra, "Kapal Terbalik di Mentawai, BMKG Ingatkan Potensi Cuaca Buruk di Pesisir dan Perairan Sumbar." Accessed: Apr. 21, 2026. [Online]. Available: https://www.kompas.id/artikel/kapal-terbalik-di-mentawai-bmkg-ingatkan-potensi-cuaca-buruk-di-pesisir-dan-perairan-sumbar

[4] H. Wang, Z. Liu, X. Wang, T. Graham, and J. Wang, "An analysis of factors affecting the severity of marine accidents," *Reliab. Eng. Syst. Saf.*, vol. 210, p. 107513, Jun. 2021, doi: 10.1016/J.RESS.2021.107513.

[5] Linda, "Transformasi Digital, BMKG Dorong Pemanfaatan AI untuk Ketangguhan Bencana Nasional - Berita - BMKG," bmkg.co.id. Accessed: Apr. 21, 2026. [Online]. Available: https://www.bmkg.go.id/berita/transformasi-digital-bmkg-dorong-pemanfaatan-ai-untuk-ketangguhan-bencana-nasional

[6] A. Scariot, A. Lo Duca, R. Lo Duca, and A. Marchetti, "MeteoChat: Semi-Automated Generation of Environmental Reports with LLMs and RAG," 2025. Accessed: Apr. 21, 2026. [Online]. Available: https://iris.cnr.it/retrieve/5b9a9e9b-19af-4577-b6f9-3e0e1c33a65a/IIT-03-2025.pdf

[7] Y. Gao, Y. Xiong, M. Wang, and H. Wang, "Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks," Jul. 2024, Accessed: Apr. 21, 2026. [Online]. Available: https://arxiv.org/pdf/2407.21059

[8] E. A. Duta-Costache, J. Aragonés-Ferrero, A. Jimeno-Morenilla, H. Mora-Mora, and J. L. Sánchez-Romero,

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.9.1.6089. SUBMITTED: APRIL 28, 2026; REVISED: JUNE 19, 2026; ACCEPTED: JUNE 24, 2026

"A Modular Microservice Architecture for Retrieval-Augmented Chatbot Systems," *2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, Jul. 2025, doi: 10.1109/ICECET63943.2025.11472574.

[9] S. M. Billah and R. J. B. R. Yusof, "Exploring the Role of RAG in Cloud-Based Monolithic Chatbots: A Systematic Review," *2025 IEEE 23rd Student Conference on Research and Development (SCOReD)*, pp. 1–6, Nov. 2025, doi: 10.1109/SCORED68498.2025.11399064.

[10] G. P. Rusum and S. Anasuri, "Vector Databases in Modern Applications: Real-Time Search, Recommendations, and Retrieval-Augmented Generation (RAG)," *International Journal of AI, BigData, Computational and Management Studies*, vol. 5, no. 4, pp. 124–136, Dec. 2024, doi: 10.63282/3050-9416.IJAIBDCMS-V5I4P113.

[11] I. A. Kamal and A. B. Cahyono, "View of Pemanfaatan Chatbot Berbasis Dialogflow dan Google Sheet Api untuk Penyimpanan Laporan Komplain Konsumen Toko Online," *Jurnal Automata*, vol. 3, no. 2, 2020, Accessed: Apr. 22, 2026. [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/24201/14033

[12] B. Kathrin, D. Georg, M. Milad, O. Simon, and S. Matthias, "A global analysis of metrics used for measuring performance in natural language processing," in *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP,* , Association for Computational Linguistics, May 2022, pp. 52–63. Accessed: Jun. 19, 2026. [Online]. Available: https://aclanthology.org/2022.nlppower-1.6.pdf

[13] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries".

[14] D. Mengliev, M. Eshkulov, V. Barakhnin, R. Abdullayev, N. Boltayev, and B. Ibragimov, "Linguistic Nuances in Text Analysis: TF-IDF Metric's Algorithm Implementation for the Karakalpak Language Recognition," *Proceedings - 2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT 2024*, pp. 19–22, 2024, doi: 10.1109/USBEREIT61901.2024.10584051.

[15] E. Filipovska *et al.*, "Evaluation of Vector Databases and LLMs in RAG-Based Multi-document Question Answering," *Communications in Computer and Information Science*, vol. 2436 CCIS, pp. 3–18, 2025, doi: 10.1007/978-3-031-86162-8_1.

[16] "Finding Word Similarity using TF-IDF and Cosine in a Term-Context Matrix from Scratch in Python | Towards Data Science." Accessed: Jun. 19, 2026. [Online]. Available: https://towardsdatascience.com/finding-word-similarity-using-tf-idf-in-a-term-context-matrix-from-scratch-in-python-e423533a407/

[17] X. Ma *et al.*, "Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 1, Dec. 2023, doi: 10.1145/3613905.3651093.

[18] "Peringatan Gelombang Tinggi," Direktorat Meteorologi Maritim. Accessed: Apr. 23, 2026. [Online]. Available: https://maritim.bmkg.go.id/cuaca/peringatan/gelombang

[19] M. Gomez-Vazquez, J. Cabot, and R. Clarisó, "Automatic Generation of Conversational Interfaces for Tabular Data Analysis," *Proceedings of the 6th Conference on ACM Conversational User Interfaces, CUI 2024*, vol. 1, Aug. 2024, doi: 10.1145/3640794.3665577.

[20] "BMKG: Perhatikan Risiko Keselamatan Pelayaran - RRI.co.id." Accessed: Apr. 23, 2026. [Online]. Available: https://rri.co.id/ambon/regional/864002/bmkg-perhatikan-risiko-keselamatan-pelayaran