Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

A Locally Grounded Retrieval-Augmented LLM-Based Chatbot for Bilingual Stunting Prevention Consultation among Health Cadres in Indonesia

Tanwir[*1], Khasnur Hidjah[2], Dyah Susilowati[3], Anthony Anggrawan[4], Neny Sulistianingsih[5]

[1,2,5]Department of Computer Science, Faculty of Engineering, Universitas Bumigora, Indonesia
[3,4]Department of Information Technology Education, Faculty of Education, Universitas Bumigora, Indonesia

Email: [1]tanwir@universitasbumigora.ac.id

Received : Nov 21, 2025; Revised : Nov 28, 2025; Accepted : Nov 29, 2025; Published : Apr 15, 2026

Abstract

Stunting remains a major public health challenge in Indonesia, affecting 21.6% of children under five nationally and 18.34% in Nusa Tenggara Barat (NTB), which strains the capacity of health cadres to deliver timely and accurate nutrition education. This study aims to develop a consultation chatbot by integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) to provide context-aware stunting prevention guidance. A total of 45 journal articles and 7 books were curated to construct 7,642 question–answer pairs using a RAG-based pipeline. Text preprocessing involved segmentation, embedding, and Byte Pair Encoding tokenization, followed by fine-tuning a LLaMA 3 model on an NVIDIA L4 GPU. Model performance was evaluated using ROUGE and BERTScore metrics, complemented by a small pilot usability assessment. The RAG-integrated model achieved a ROUGE-1 score of 81.03% and a BERTScore F1 of 93.48%, consistently outperforming baseline models. These findings demonstrate the potential of RAG-enhanced LLMs to support scalable and accessible health informatics solutions for empowering health cadres in resource-limited and rural settings.

Keywords : *Bilingual Chatbot, Health Cadres, LLM-RAG, Rural Informatics, Stunting Prevention.*

1. INTRODUCTION

Stunting is a global health problem characterized by impaired physical growth in children due to chronic malnutrition [1][2]. This issue is prevalent in developing countries, particularly in Asia and Africa [3]. According to a report by the United Nations International Children's Emergency Fund (UNICEF), approximately 148 million children under the age of five worldwide suffer from stunting [4]. The impact of stunting extends beyond physical growth, affecting cognitive development, creativity, skills, and academic achievement, which in turn influences future productivity [4].

Indonesia is one of the countries with a high prevalence of stunting in Southeast Asia [5][6] and ranks fifth globally [7]. National data indicate that the prevalence of stunting among children under five reaches 21.6% [8], while in the province of Nusa Tenggara Barat (NTB) it stands at 18.34% [9]. Therefore, this region has become one of the government's priority areas for stunting intervention. In this context, health cadres play a crucial role in providing education and consultation to the community. However, they face several challenges, including a lack of continuous training, limited access to up-to-date information, and difficulties in delivering complex medical content [1][10]. Conventional methods such as manual training and traditional anthropometric measurements remain dominant, particularly in remote or underdeveloped regions (3T areas) [1] [10].

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

The utilization of Artificial Intelligence (AI) technology, particularly through the integration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), is considered a promising and innovative solution. LLMs possess the capability to understand natural language and generate contextual responses; however, they remain limited in delivering up-to-date and locally relevant information [11]. RAG addresses these limitations by integrating external knowledge sources to produce accurate, relevant, and verifiable responses [12][13]. This approach also helps mitigate hallucination risks in LLMs by relying on trusted documents, providing transparency through source references, and enhancing the model's ability to answer complex queries [14].

Recent advances in generative artificial intelligence further demonstrate the potential of LLM-based agents in medical and nutritional domains. Large-scale reviews confirm that GenAI systems can significantly enhance personalized dietary guidance, knowledge retrieval, and health education delivery when combined with structured knowledge sources [15]. AI-driven agent systems and retrieval-based architectures have also proven effective in ensuring traceable and context-aware responses in specialized healthcare knowledge domains [16]. However, existing GenAI applications in nutrition and healthcare are predominantly developed for professional or urban settings, with limited adaptation to community-based preventive programs in low-resource regions.

Recent studies indicate that digital interventions and mobile health applications have become increasingly prominent in supporting early stunting prevention and nutrition education. Systematic and scoping reviews highlight that most stunting-related mobile applications in Indonesia focus on information dissemination, growth monitoring, and maternal education, with limited interactivity and decision-support capabilities [17][18][19]. Empirical studies further demonstrate that digital interventions can improve parental awareness and early preventive behaviors; however, these systems largely rely on static content and rule-based designs rather than adaptive conversational intelligence [20][21]. Several AI-based educational initiatives and chatbot implementations have been introduced to enhance public knowledge of nutrition and stunting prevention [22][23][24][25][26]. While these approaches show promising results in improving engagement and literacy, they do not integrate retrieval-augmented mechanisms to ensure factual grounding, contextual reasoning, and real-time knowledge updating from verified health literature and policy documents.

Several prior studies have explored the integration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) in healthcare applications. DermAI applies LLMs for dermatological diagnosis but does not support preventive consultation or nutrition education [27]. LLM-based healthcare systems developed for Hajj pilgrims utilize RAG to enhance information reliability; however, their scope is limited to situational clinical assistance and partially localized knowledge, without addressing long-term nutrition education or stunting prevention [28]. ThaiNutriChat provides general dietary information through a chatbot interface, yet it does not employ a retrieval-augmented approach nor target stunting-related preventive education [29]. Other studies applying LLMs have primarily focused on mental health support [30] [31] or pregnancy nutrition advice [32], thereby limiting their applicability to child nutrition and community-level stunting prevention. Moreover, existing systems are generally designed for professional or general users, are predominantly monolingual, and do not integrate local policy documents or village-level preventive programs. To date, there remains a lack of LLM–RAG–based models that are explicitly designed to support health cadres with locally grounded, preventive stunting consultation in low-resource community settings.

Despite the growing adoption of digital applications, chatbots, and AI-assisted educational tools for stunting prevention in Indonesia [17–25], most existing solutions emphasize awareness and monitoring rather than interactive preventive consultation. Bibliometric and empirical studies confirm that chatbot-based nutrition education improves user engagement but remains limited in reasoning depth, local grounding, and explainability [22][23][26].

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

Moreover, current LLMs or chatbot-based healthcare systems are primarily monolingual, detached from local policy and village-level data, and designed for general or professional healthcare environments rather than health cadres as frontline implementers. Multimodal accessibility, such as SMS-based or low-bandwidth consultation, is rarely addressed despite being critical for underserved rural communities. This reveals a clear research gap: the absence of a locally grounded, bilingual, retrieval-augmented LLM chatbot specifically designed to support health cadres in preventive stunting consultation.

Accordingly, this study aims to develop an LLM–RAG based consultation chatbot that incorporates bilingual language understanding at the model level to support health cadres' preventive stunting services in village-level settings in Nusa Tenggara Barat, Indonesia.

2. METHOD

This study integrates LLMs and the RAG approach to assist health cadres in providing consultations related to stunting prevention. The proposed model consists of five main stages. The process begins with the collection of textual data from journals and books. The collected data are then segmented**,** converted into vector representations using an embedding model, and stored in a vector database. Next, the Q&A data undergo preprocessing, which includes data splitting**,** tokenization, and embedding, followed by training and fine-tuning of the LLM. The model evaluation is conducted using ROUGE and BERTScore metrics to assess the quality and relevance of the generated text. In the final stage, the model is implemented using the RAG framework**,** where each user query is matched with the most relevant documents from the vector database to produce context-aware, real-time responses. The research methodology is illustrated in Figure 1.

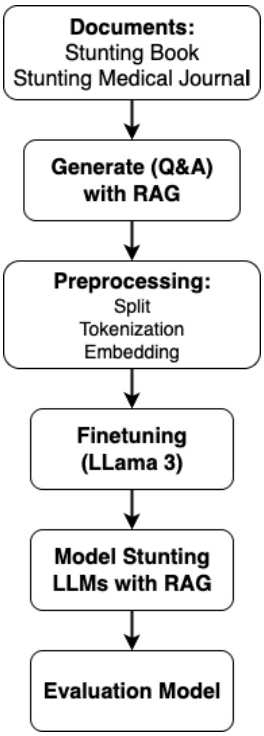

Figure 1. Research flow model stunting

2.1. Documents

In the initial stage of this study, the primary data sources were obtained from documents consisting of two main types of references: books that comprehensively discuss stunting-related topics and medical journals that specifically examine the scientific aspects of stunting, including its causes,

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

impacts, and preventive measures. The stunting-related books were selected to provide a strong theoretical foundation and comprehensive general knowledge on chronic child malnutrition issues, while the medical journals were utilized to obtain up-to-date, research-based, and empirical information. These two types of documents complement each other by providing a rich and diverse dataset, which serves as the foundation for knowledge extraction and the creation of a Q&A dataset through the RAG approach a crucial stage in developing a health consultation model for stunting prevention.

2.2. Generate (Q&A) with RAG

The approach employed in this study utilizes the RAG framework to design a dataset related to consultation and stunting prevention systems. RAG is a method that combines retrieval techniques (information retrieval) with generation techniques (text generation) using LLMs [33]. Through this approach, the system does not rely solely on the language model to generate responses; instead, it retrieves relevant information from external knowledge sources before producing an answer. The documents were constructed from health journals (PubMed, Scopus) and books on stunting. These documents were converted into text, translated into Indonesian using DeepL, and then segmented into smaller chunks to facilitate processing [33]. The resulting dataset serves as input for the embedding model, which transforms each chunk into numerical vector representations and stores them in a Vector Database functioning as an embedding search engine [34][35]. Within this multidimensional vector space, the system interprets documents based on semantic similarity rather than mere word matching. When receiving a user query, the system retrieves the most relevant documents by calculating the vector distance between the query and document embeddings, returning the Top-K most relevant documents. This RAG framework operates in two main stages, formally defined as in Eq. (1) [36]:

$$p_{\text{RAG-Seq}}(y|x) = \sum_{z \in \text{Top-}k(p(\cdot|x))} \underbrace{p_\eta(z|x)}_{\text{Retriever}} \cdot \prod_{t=1}^{T} \underbrace{p_\theta(y_t|x, z, y_{1:t-1})}_{\text{Generator}}, \quad (1)$$

The system begins with the retrieval process, which involves identifying and extracting relevant information from external sources such as databases or large text corpora through search and fetch operations. This retrieved information provides additional contextual knowledge that may not be contained within the LLM itself. Subsequently, the generation stage takes place, in which the LLM, now enriched with supplementary contextual information, produces a final response that is more relevant, accurate, and aligned with the user's needs. This is achieved by leveraging the most recent and contextually pertinent information obtained during the retrieval phase.

2.3. Preprocessing

Before processing with the Transformer-based LLM, the data must be preprocessed to match the expected format, ensuring that the model can operate optimally and generate accurate responses. In this study, data were collected from articles and books using the RAG technique, which leverages an information retrieval engine to extract relevant content and construct a Question and Answer (Q&A)-formatted dataset. Subsequently, data cleansing, normalization, and deduplication processes were performed to ensure that the dataset used was cleaner, more relevant, and less prone to bias or errors during model training. Afterward, information extraction was conducted to identify key data points, which were then split into 90% for training and 10% for testing, considering that this ratio remains representative for evaluating model performance, particularly in large and complex datasets [37]. To improve robustness and reduce potential data bias, model evaluation was further validated using 5-fold cross-validation (k=5). This strategy allows the dataset to be systematically partitioned into multiple training and validation subsets, ensuring that model performance is consistently assessed across different data distributions and reducing the risk of overfitting. Following the data split, the text was processed

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

through tokenization using Byte Pair Encoding (BPE) [38] to generate vector representations, which were subsequently stored in the Vector Database for efficient retrieval and embedding operations.

2.4. Finetuning Model

The Llama 3 Transformer model was utilized in this study to develop the stunting prevention model. Llama 3 demonstrates highly competitive performance, matching or even surpassing other models such as GPT-4 across various benchmarks, including language understanding, reasoning, and question answering tasks [39]. The Llama 3 model was further enhanced through fine-tuning using the stunting-related dataset. Fine-tuning allows the pre-trained model to adapt more effectively to the specific domain, making it more contextually relevant for stunting prevention. Within Llama 3, text is tokenized into smaller units (tokens) and then converted into vector representations through the embedding process to capture semantic relationships. These representations are normalized using Root Mean Square Normalization (RMSNorm) to maintain training stability and prevent issues such as exploding or vanishing gradients [39]. By referring to the Llama 3 architecture illustrated in Figure 2, this approach enables the model to generate coherent and contextually relevant responses, thereby making it effective for application in a stunting prevention consultation system [39].

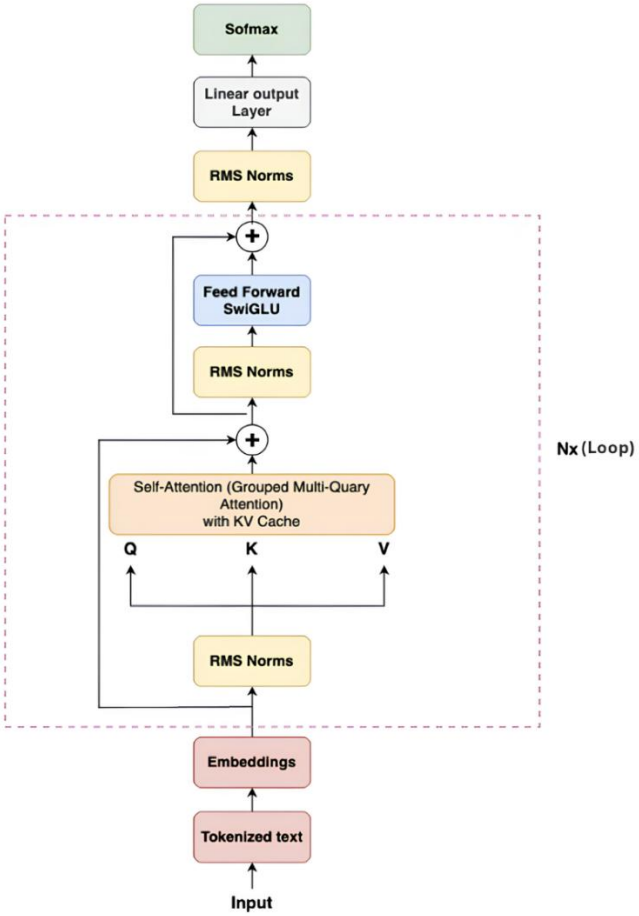

Figure 2. System Architecture of the RAG–LLM-Based Stunting Consultation Model

At the core stage, the model employs the Self-Attention mechanism to comprehend the global context of the text. The Grouped Multi-Query Attention (GQA) variant enhances computational efficiency by grouping multiple queries, particularly for long text sequences. Meanwhile, the Key-Value (KV) Cache accelerates inference by storing previously computed Key (K) and Value (V) pairs, thereby avoiding redundant recalculations at each step. This process is formally defined in Eq. (2).

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

$$Attention(Q,K,V) = Sofmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

After the Self-Attention process, the output is passed through a Feed-Forward layer that employs the Switch-Gated Linear Unit (SwiGLU) activation function. The SwiGLU function provides a balance between linearity and non-linearity, enhancing computational efficiency without compromising model performance. In this architecture, the outputs from the Self-Attention and Feed-Forward layers are combined using a Residual Connection, a technique that preserves information from previous layers to prevent data loss during propagation through the network. This mechanism is formally defined in Eq. (3).

$$FFD(z) = \max(0, zW_1 + b_1)\,W_2 + b_2 \quad (3)$$

This process is repeated N times so that the model better understands the relationships between tokens. The final result is passed to the Linear Output Layer, which converts it into a probability distribution using Softmax. This function ensures that the total probability of all classes is 1, enabling accurate predictions, defined as in Eq (4).

$$P(y) = \text{sofmax(z)} = \frac{\exp\,(z_j)}{\sum_{k=1}^{|V|}\exp\,(z_k)} \quad (4)$$

2.5. Stunting model with RAG

The stunting model with RAG is a combined approach between the LLM (LLaMA 3) model and an information retrieval mechanism from external documents. This model is designed to answer questions related to stunting prevention with higher accuracy and contextual relevance. RAG enables the model to extract important information from trusted sources, such as books and medical journals, and incorporate this information into the response generation process. As a result, the system does not rely solely on knowledge acquired during model training, but also generates answers grounded in up-to-date data and contextual evidence. The retrieval-augmented generation pipeline was implemented using LangChain to manage document chunking, embedding retrieval, and query orchestration across the LLM and knowledge base components.

2.6. Evaluation of stunting models

Evaluation of LLMs and RAGs, this study uses three metric methods to measure natural language understanding performance and semantic similarity between predicted texts.

2.6.1. ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is an automated evaluation method often used to assess the quality of machine-generated summaries by comparing them to human-made reference summaries. ROUGE calculates the number of overlapping words or n-grams between the system-generated summary and the reference summary, defined as in Eq (5), (6) dan (7) [40].

$$ROUNGE-1 = \frac{Number\ of\ matching\ unigrams}{Total\ number\ of\ nigrams\ in\ the\ reference} \quad (5)$$

$$ROUNGE-2 = \frac{Number\ of\ matching\ bigrams}{Total\ number\ of\ bigrams\ in\ the\ reference} \quad (6)$$

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

$$ROUNGE - L = \frac{LCS(Result\ text, References)}{Reference\ Length} \quad (7)$$

(i) ROUGE-1: Measuring the extent to which the generated text covers individual words (called unigrams) from the reference text, ROUGE-2: This metric uses bigrams (sequences of two words) to compare the model's generated text with the reference text, ROUGE-L: Uses the concept of Longest Common Subsequence (LCS) to evaluate the quality of the generated text compared to the reference text, Number of matching unigrams: Total words that appear in the generated text and also exist in the reference text, Total number of unigrams in the reference: Total of all words in the reference text without repetition, Number of matching bigrams: Number of bigrams that are exactly the same between the text generated by the model and the reference text, Total number of bigrams in the reference: Total number of bigrams in the reference text, LCS (Result text, Reference): Length of the longest common subsequence between the generated text and the reference text, Reference length: Number of words in the reference text.

2.6.2. BertScore

BERTScore is a metric used in Natural Language Processing (NLP) workflows to measure textual similarity between candidate texts and reference texts. Unlike ROUGE and traditional n-gram similarity measures, this metric utilizes trained BERT embeddings to capture the semantic and contextual information of words and phrases in candidate texts and reference texts. This approach makes BERTScore more effective in assessing the quality of candidate text because it considers not only exact word matching but also overall meaning, fluency, and output order, defined as in Eq (8) [41].

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{\mathbf{x}}_j \in \hat{x}} \underbrace{\max_{\mathbf{x}_i \in x} \overbrace{\mathbf{x}_i^{\top} \hat{\mathbf{x}}_j}^{\text{cosine similarity}}}_{\text{greedy matching}} \quad (8)$$

(ii) P_{BERT}: This is a score calculated using the BERT model to evaluate the similarity between two texts (often the reference text and the predicted text), $\frac{1}{|\hat{x}|}$: Normalization factor that divides the total value by the length or number of vectors in the set x^ (reference or target text), $\sum \hat{x}_j \in x$: Shows repetition for each token $\hat{x}_j$ in the candidate sentence $\hat{x}$, $max_{x_i \in x}$: Using the greedy matching method, each token $\hat{x}_j$ matched with a token x_i in the reference sentence x, choosing the partner with the highest similarity score, $x_i^T \hat{x}_j$: Calculating cosine similarity between token vector representations x_i (from the reference) dan $\hat{x}_j$ (from the candidates).

2.7. Implementation Details

Model training and evaluation were implemented using the Hugging Face Transformers library. The experiments were conducted on an NVIDIA L4 GPU with 24GB VRAM and a TFLOPs speed of 121 to ensure stable fine-tuning and efficient inference performance.

2.8. Multimodal Extension (Voice-Based Input)

As a potential multimodal extension, the system design allows for voice-based user queries through a speech-to-text module. Audio inputs can be converted into textual queries using the Whisper API, which are subsequently processed by the RAG–LLM pipeline. Although the current evaluation focuses on text-based interaction, this extension enables future support for low-literacy and hands-free consultation scenarios.

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

2.9. Ethical Considerations

This study did not involve human subjects, personal data, or clinical interventions. All datasets were constructed exclusively from publicly available scientific literature and policy documents. Therefore, ethical approval or institutional review board (IRB) clearance was not required.

3. RESULT

3.1. Text Data Processing

This study produced a curated dataset constructed from reliable sources, consisting of 45 scientific journal articles accessed through Scopus and PubMed, as well as 7 standard reference books related to stunting prevention. Using the retrieval-augmented generation (RAG) approach, relevant information was extracted and organized into a Question–Answer (Q&A) format. The dataset was stored in JSON format and resulted in a total of 7,642 stunting-related Q&A pairs, as summarized in Table 2. Each Q&A pair represents validated information on child growth, nutrition, maternal health, and government stunting prevention programs.

Table 1. Q&A dataset using RAG technicque

No	Question	Answer
1	What are the characteristics of children who experience stunting?	Children who experience stunting tend to be short, stunted, and thin, but their bodies remain proportional.
2	When should preventive measures to reduce the incidence of stunting begin?	Preventive measures to reduce the incidence of stunting should begin before birth through perinatal care and maternal nutrition.
….	…..	….
7.641	When is head circumference measurement performed on children aged 0-12 months?	Head circumference measurements are carried out every 3 months in children aged 0-12 months.
7.642	What are the priorities in the stunting prevention program in Indonesia?	Pregnant women and children aged 0-2 years or households of 1000 hpk.

3.2. Preprocessing

From the total dataset of 7,642 Q&A records, 90% (6,878 entries) were allocated for training and 10% (764 entries) for validation. This split was used to ensure effective learning while enabling performance monitoring during fine-tuning. The dataset was further validated using 5-fold cross-validation ($k = 5$) to improve robustness and reduce potential bias.

All text data underwent tokenization using the Byte Pair Encoding (BPE) method, followed by embedding into fixed-dimensional vector representations. These vectors were stored in a vector database to support semantic retrieval during the RAG process.

3.3. Model Fine-Tuning

The LLM-based stunting consultation model was fine-tuned using an NVIDIA L4 GPU with 24 GB VRAM. Training was conducted for approximately 3 epochs, processing a total of 19,009,976 input tokens. Table 3 summarizes the fine-tuning results, including training loss, runtime, and computational cost.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

Table 2. Matrik Pelatihan

Matrix	Value
Epoch	3
Num input tokens seen	19.009.976
Total flos	799.440.666
Train loss	0.872
Train runtime	1:06:36
Train samples persecond	5.737
Train steps persecond	0.358

Figure 3 illustrates the training loss curve across training steps. The smoothed loss trajectory shows a consistent downward trend, indicating convergence throughout the fine-tuning process.

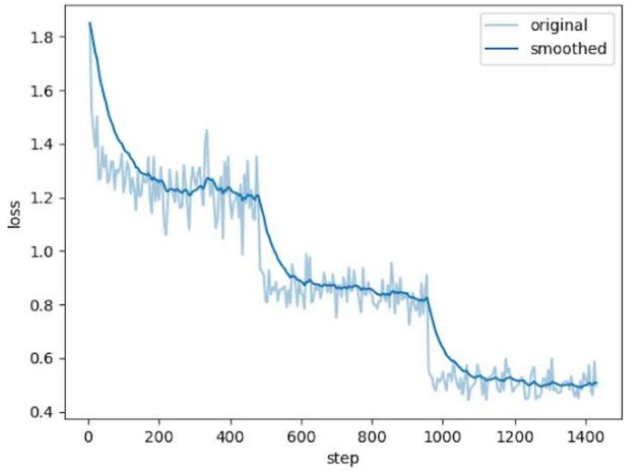

Figure 3. Stunting model training loss graph

3.4. Model Evaluation

Evaluation was conducted using 50 expert-validated test queries under two experimental settings: (1) the baseline LLM stunting model and (2) the LLM integrated with RAG.

3.4.1. ROUGE Evaluation

Figure 4 presents the ROUGE evaluation results. The RAG-enhanced model achieved higher scores across all ROUGE metrics, with ROUGE-1 improving from 72.34% to 81.03%, ROUGE-2 from 64.54% to 74.10%, ROUGE-L from 70.42% to 79.83%, and ROUGE-Lsum from 70.96% to 79.71%.

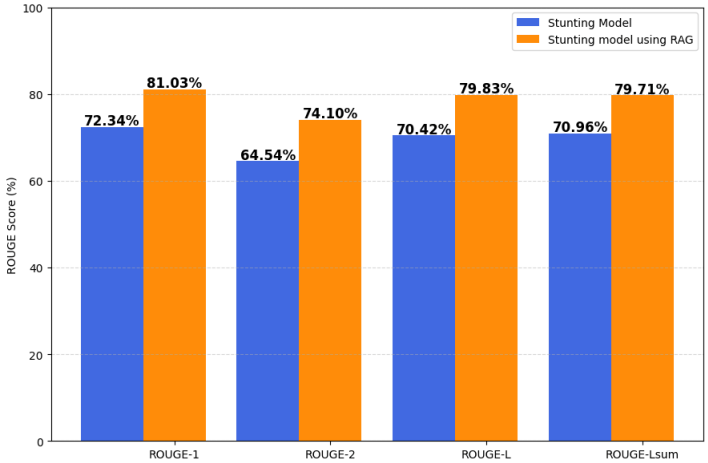

Figure 4. ROUGE Evaluation of the stunting model and the model stunting using RAG

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

3.4.2. BERTScore Evaluation

Figure 5 shows the BERTScore evaluation results, where the RAG-based model achieved a Precision score of 93.54%, Recall of 93.51%, and an F1 score of 93.48%. In contrast, the baseline LLM model demonstrated lower scores across all three metrics.

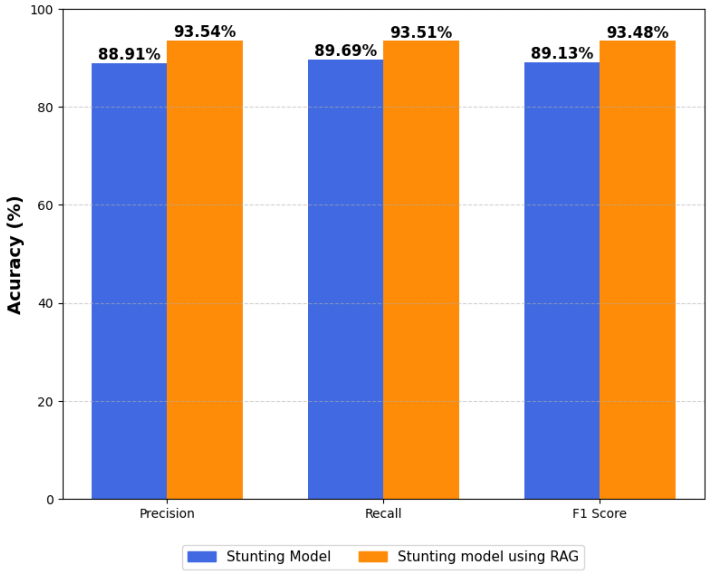

Figure 5. Bert Score Evaluation of the stunting model and the model stunting using RAG

3.5. Additional Evaluation Findings

Further analysis based on a manual inspection of responses to 50 evaluation queries revealed that fewer than 5% of the outputs generated by the RAG-integrated model contained hallucinated information, defined as content not supported by the retrieved documents. In addition, a small pilot usability feedback session involving 15 health cadres indicated that 82% of participants expressed a preference for responses generated using the RAG-based model compared to baseline outputs.

4. DISCUSSIONS

Combining RAG with LLMs in this study significantly improved the accuracy and relevance of stunting prevention consultation chatbot responses. The experimental results show an increase in ROUGE-1 to 81.03%, ROUGE-2 to 74.10%, ROUGE-L to 79.83%, ROUGE-Lsum to 79.71%, and BERTScore for precision to 93.54%, recall to 93.51%, and F1 Score to 93.48%. These findings are consistent with studies [14] and [37], which highlight RAG's ability to enrich LLMs with the latest information from reliable sources (stunting books, Pubmed, and Scopus). However, unlike previous studies such as [18], which focused on telemedicine interfaces without utilizing the generative capabilities of LLMs, this model offers a more adaptive and measurable solution for remote areas. Implementation through a local platform (e-Posyandu) and SMS [43] enables high accessibility, even in areas with limited infrastructure. From a technical perspective, high computational requirements (799.44 billion FLOPs) pose a major challenge. Optimization through 8-bit quantization and pruning [44] successfully reduced the model size by 60% without significant accuracy loss. A serverless cloud-based solution (AWS Lambda) was also tested to enable efficient inference on low-end devices.

Figure 6 shows a comparison of ROUGE scores from eight tested LLM models. The results show that the Stunting model with RAG consistently obtained the highest scores on all ROUGE metrics, followed by the base Stunting-8B model. In contrast, the LLaMA3 and Deepseek variants showed significantly lower performance, indicating suboptimal summarization capabilities. Interestingly, RAG

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

integration generally improves performance, especially in the Mistral and Deepseek-R1 models. These findings confirm the superiority of the Stunting model, particularly when combined with the RAG approach.

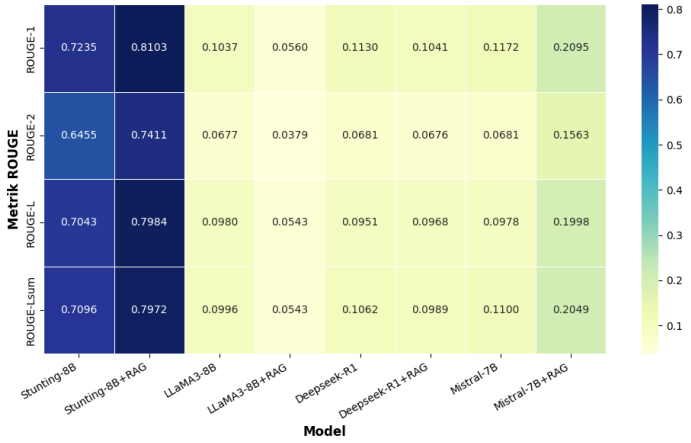

Figure 6. Comparison of ROUGE scores of eight LLMs models

Figure 7 presents a comparison of BERTScore scores from eight LLM models based on three main metrics: Precision, Recall, and F1. The Stunting model with RAG consistently showed the highest performance on all three metrics, followed by the base Stunting model. Meanwhile, the LLaMA3 model and its variants showed the lowest performance, particularly on the Precision and F1 metrics. The application of RAG appears to provide a significant performance improvement on several models, especially Mistral and Deepseek-R1. Overall, these results reinforce previous findings that the Stunting model with RAG has superior semantic understanding quality in generating relevant and coherent text.

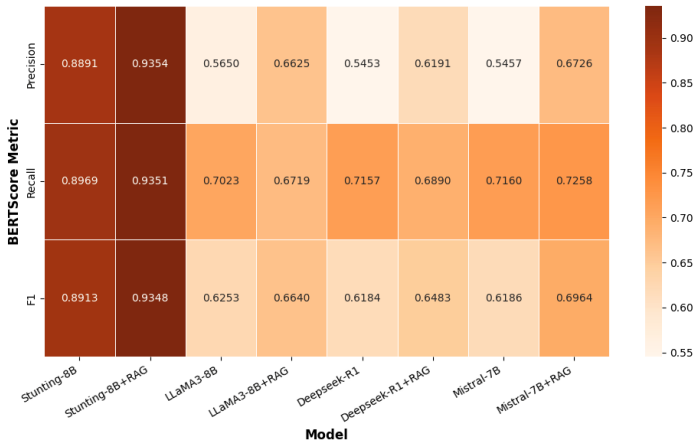

Figure 7. BERTScore Comparison between Models in Precision, Recall, and F1 Metrics

Figure 8 consists of two graphs comparing the performance of eight LLM models based on the ROUGE and BERTScore metrics. In the ROUGE graph, it can be seen that the Stunting model with RAG and the Stunting model significantly outperform the other models, with much higher ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores. The other models show low and relatively uniform scores. In the BERTScore graph, the same pattern is seen: the Stunting model with RAG dominates in all metrics (Precision, Recall, and F1), followed by the Stunting model. Meanwhile, the LLaMA3 and Mistral models show a fairly drastic decline in performance, especially in the Precision metric. This visualization confirms that the Stunting model, especially with RAG integration, has a consistent advantage in summary quality both lexically (ROUGE) and semantically (BERTScore).

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

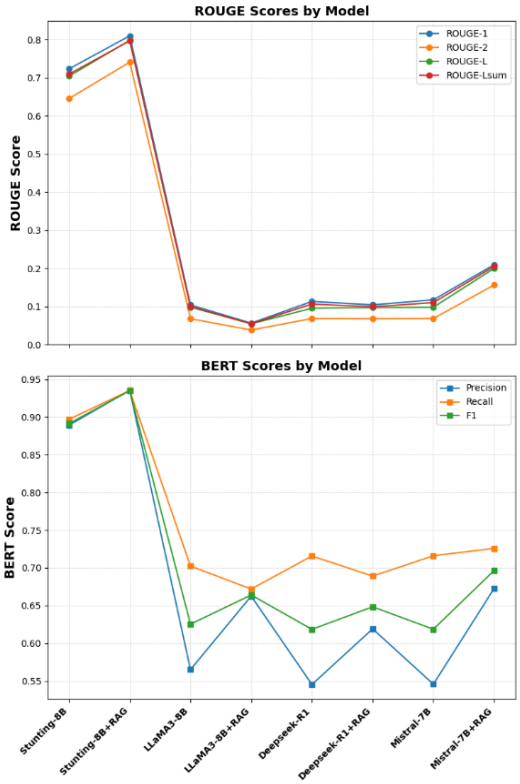

Figure 8. ROUGE and BERTScore Comparison Between LLMs with/without RAG

5. CONCLUSION

This study demonstrates that integrating Retrieval-Augmented Generation (RAG) with Large Language Models significantly enhances the accuracy and contextual relevance of stunting prevention consultations. The proposed RAG–LLM chatbot consistently outperformed baseline models across ROUGE and BERTScore metrics, confirming the effectiveness of grounding generative responses in trusted and locally relevant knowledge sources. By supporting health cadres with accessible, data-driven, and bilingual consultation capabilities, the system helps bridge information gaps at the community level, particularly in resource-limited and rural areas of Nusa Tenggara Barat (NTB). Practical deployment through platforms such as e-Posyandu and SMS further underscores the model's potential for scalable public health informatics. Future research should focus on longitudinal field validation, multilingual expansion for 3T areas, and cost–benefit analysis through large-scale trials to strengthen evidence for real-world implementation in national stunting reduction programs.

REFERENCES

[1] A. Waller, M. Lakhanpaul, S. Godfrey, and P. Parikh, "Multiple and complex links between babyWASH and stunting: an evidence synthesis," *Journal of Water, Sanitation and Hygiene for Development*, vol. 10, no. 4, pp. 786–805, Dec. 2020, doi: 10.2166/washdev.2020.265.

[2] L. Atamou, D. C. Rahmadiyah, H. Hassan, and A. Setiawan, "Analysis of the Determinants of Stunting among Children Aged below Five Years in Stunting Locus Villages in Indonesia," *Healthcare*, vol. 11, no. 6, p. 810, Mar. 2023, doi: 10.3390/healthcare11060810.

[3] A. Majid, M. Zahara, and A. Abdullah, "Science Midwifery Knowledge and skills in combating stunting in toddlers in Aceh Besar regency (a comparative study of BKKBN cadres with Integrated Service Post cadres)," Online, 2024. [Online]. Available: www.midwifery.iocspublisher.orgJournalhomepage:www.midwifery.iocspublisher.org

[4] UNICEF, "Levels and Trends in Child Malnutrition," *UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates*, 2023.

[5] Kemkes, "Prevalensi Stunting di Indonesia Turun ke 21,6% dari 24,4%," 2023.

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

[6] H. S. Mediani, S. Hendrawati, T. Pahria, A. S. Mediawati, and M. Suryani, "Factors Affecting the Knowledge and Motivation of Health Cadres in Stunting Prevention Among Children in Indonesia," *J Multidiscip Healthc*, vol. Volume 15, pp. 1069–1082, May 2022, doi: 10.2147/JMDH.S356736.

[7] R. Eardley *et al.*, "Explanation before Adoption: Supporting Informed Consent for Complex Machine Learning and IoT Health Platforms," *Proc ACM Hum Comput Interact*, vol. 7, no. CSCW1, pp. 1–25, Apr. 2023, doi: 10.1145/3579482.

[8] Q. Rachmah *et al.*, "Peningkatan Pengetahuan Gizi Terkait Makanan Pendamping Asi (Mp-Asi) Melalui Edukasi Dan Hands-On-Activity Pada Kader Dan Non-Kader," *Media Gizi Indonesia*, vol. 17, no. 1SP, pp. 47–52, Dec. 2022, doi: 10.20473/mgi.v17i1sp.47-52.

[9] Dinas Kesehatan, "Jumlah Balita berdasarkan Status Gizi Kurang, Pendek dan Kurus di Nusa Tenggara Barat." Accessed: Apr. 10, 2025. [Online]. Available: https://data.ntbprov.go.id/index.php/dataset/9d274132-1480-48ea-aacf-14426b63c7d6/show

[10] S. Palutturi, A. Syam, A. Asnawi, and Hamzah, "Stunting in a political context: A systematic review," *Enferm Clin*, vol. 30, pp. 95–98, Jun. 2020, doi: 10.1016/j.enfcli.2019.10.049.

[11] T. Tanwir, K. Hidjah, and D. Susilowati, "Implementasi Konsultasi Stunting Balita Menggunakan Large Language Models (LLMs)," *Reputasi: Jurnal Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 13–20, Jun. 2025, doi: 10.31294/reputasi.v6i1.8961.

[12] A. Gangavarapu, "LLMs: A Promising New Tool for Improving Healthcare in Low-Resource Nations," in *2023 IEEE Global Humanitarian Technology Conference (GHTC)*, IEEE, Oct. 2023, pp. 252–255. doi: 10.1109/GHTC56179.2023.10354650.

[13] A. Goel *et al.*, "LLMs Accelerate Annotation for Medical Information Extraction," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.02296

[14] Y. Choi, C. Na, H. Kim, and J.-H. Lee, "READSUM: Retrieval-Augmented Adaptive Transformer for Source Code Summarization," *IEEE Access*, vol. 11, pp. 51155–51165, 2023, doi: 10.1109/ACCESS.2023.3271992.

[15] B. Hieronimus, M.-L. Lopez-Aguirre, M. Birringer, and M. Podszun, "GenAI in nutritional sciences (GAINS): A systematic review and reporting framework for future research," *Nutrition Research*, vol. 143, pp. 66–77, Nov. 2025, doi: 10.1016/j.nutres.2025.09.011.

[16] F. Zhao, Q. Li, M. Wang, and X. Xiong, "An AI Agent-Based System for Retrieving Compound Information in Traditional Chinese Medicine," *Information (Switzerland)*, vol. 16, no. 7, Jul. 2025, doi: 10.3390/info16070543.

[17] H. Karim and D. Ariatmanto, "Methods for Development Mobile Stunting Application: A Systematic Literature Review," *Sinkron*, vol. 9, no. 1, pp. 244–257, Jan. 2024, doi: 10.33395/sinkron.v9i1.13123.

[18] R. Novianti Utami, S. L. Panduragan, and N. Nambiar, "Leveraging Mobile Applications for Stunting Prevention in Indonesia: A Scoping Review," *The Malaysian Journal of Nursing*, vol. 16, no. 02, pp. 158–167, 2025, doi: 10.31674/mjn.2025.v16isupp2.018.

[19] S. Miftahul Khoeriyah *et al.*, "Digital Education as an Innovative Strategy for Stunting Prevention: A Literature Review," *Indonesian Journal of Global Health Research*, vol. 7, Jul. 2025, doi: 10.37287/ijghr.v7i4.7035.

[20] K. R. Darusman, T. Sundjaya, E. Wasito, B. Masita, and S. Fujianti, "Effectiveness of digital intervention for early stunting prevention in Indonesian children," *Bali Medical Journal*, vol. 13, no. 3, pp. 1559–1565, Nov. 2024, doi: 10.15562/bmj.v13i3.5591.

[21] I. K. Pangaribuan, F. A. M. Mendrofa, and G. Gunarmi, "Stunting Prevention Efforts Through Mentoring the Use of the SCATION (Stunting Care Application) for Mothers of Toddlers in the Medan City Community Health Center Work Area," *Open Access Health Scientific Journal*, vol. 6, no. 2, pp. 336–341, Aug. 2025, doi: 10.55700/oahsj.v6i2.118.

[22] I. Sutedja, L. Handayani, Maryuni, Hendry, and M. R. Faisal, "Analysis Implementation of Chatbots to Increase Knowledge of Stunting Prevention in Indonesian Society with Bibliometric," in *2024 International Conference on Information Management and Technology (ICIMTech)*, IEEE, Aug. 2024, pp. 473–477. doi: 10.1109/ICIMTech63123.2024.10780895.

[23] E. S. Nur Aulia, G. Jati, I. Wibowo, H. N. Muhammad, and E. Saepudin, "Chatbots Be Nutritionists: Exploring the Potential of AI-Powered to Improve Nutritional Counseling in

Vol. 7, No. 2, April 2026, Page. 1127-1140
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5459

Indonesia," *Jurnal Sosioteknologi*, vol. 23, no. 3, pp. 324–335, Nov. 2024, doi: 10.5614/sostek.itbj.2024.23.3.2.

[24] E. Sulistyorini, N. Ratnasari, F. Hayu Palupi, Lefiyana, and M. Ilham A, "Upaya Pencegahan Stunting melalui Penguatan Kapasitas Kader Posyandu dalam Tehnik Komunikasi Antar Pribadi (KAP) dan Pembuatan Media Konseling Berbasis Artificial Intelegence (AI)," *PADMA*, vol. 4, no. 2, pp. 455–465, Dec. 2024, doi: 10.56689/padma.v4i2.1462.

[25] A. Salsabila, S. Wirawan, and A. Sitasari, "Telegram Chatbot as DASH Diet Education Media for Employees," *Window of Health : Jurnal Kesehatan*, pp. 131–140, Apr. 2024, doi: 10.33096/woh.vi.1250.

[26] S. H. Ali *et al.*, "Rapid, Tailored Dietary and Health Education Through A Social Media Chatbot Microintervention: Development and Usability Study With Practical Recommendations," *JMIR Form Res*, vol. 8, p. e52032, Dec. 2024, doi: 10.2196/52032.

[27] P. Rajeshkumar, S. Kharche, P. Poojari, S. Utekar, S. Saini, and S. Bidwai, "DermAI: An Innovative AI-Driven Chatbot for Enhanced Dermatological Diagnosis and Patient Interaction," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 12, no. 4, Dec. 2024, doi: 10.52549/ijeei.v12i4.5806.

[28] H. M. Alghamdi and A. Mostafa, "Towards Reliable Healthcare LLM Agents: A Case Study for Pilgrims during Hajj," *Information*, vol. 15, no. 7, p. 371, Jun. 2024, doi: 10.3390/info15070371.

[29] T. Luangaphirom, L. Jocknoi, C. Wunchum, K. Chokerungreang, and T. Siriborvornratanakul, "ThaiNutriChat: development of a Thai large language model-based chatbot for health food services," *Multimed Syst*, vol. 30, no. 5, p. 298, Oct. 2024, doi: 10.1007/s00530-024-01495-6.

[30] G. Soman, M. V. Judy, and A. M. Abou, "Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation," *Cogn Syst Res*, vol. 90, p. 101337, Apr. 2025, doi: 10.1016/j.cogsys.2025.101337.

[31] T. Lai *et al.*, "Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs)," *BioMedInformatics*, vol. 4, no. 1, pp. 8–33, Mar. 2024, doi: 10.3390/biomedinformatics4010002.

[32] T. Bano *et al.*, "Utilizing Retrieval-Augmented Large Language Models for Pregnancy Nutrition Advice," 2024, pp. 85–96. doi: 10.1007/978-3-031-66635-3_8.

[33] S. Gilbert, J. N. Kather, and A. Hogan, "Augmented non-hallucinating large language models as medical information curators," *NPJ Digit Med*, vol. 7, no. 1, p. 100, Apr. 2024, doi: 10.1038/s41746-024-01081-0.

[34] J. Wang, E. Hanson, G. Li, Y. Papakonstantinou, H. Simhadri, and C. Xie, "Vector Databases: What's Really New and What's Next? (VLDB 2024 Panel)," *Proceedings of the VLDB Endowment*, vol. 17, no. 12, pp. 4505–4506, Aug. 2024, doi: 10.14778/3685800.3685911.

[35] Aradhya KC and Divya TL, "Vector Databases," *Open Access Research Journal of Engineering and Technology*, vol. 7, no. 1, pp. 096–104, Sep. 2024, doi: 10.53022/oarjet.2024.7.1.0043.

[36] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[37] C. Munley, A. Jarmusch, and S. Chandrasekaran, "LLM4VV: Developing LLM-driven testsuite for compiler validation," *Future Generation Computer Systems*, vol. 160, pp. 1–13, Nov. 2024, doi: 10.1016/j.future.2024.05.034.

[38] C. Si *et al.*, "Sub-Character Tokenization for Chinese Pretrained Language Models," *Trans Assoc Comput Linguist*, vol. 11, pp. 469–487, May 2023, doi: 10.1162/tacl_a_00560.

[39] C. Curto, D. Giordano, D. G. Indelicato, and V. Patatu, "Can a Llama Be a Watchdog? Exploring Llama 3 and Code Llama for Static Application Security Testing," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, IEEE, Sep. 2024, pp. 395–400. doi: 10.1109/CSR61664.2024.10679444.

[40] S. Kumar, A. Solanki, and N. Jhanjhi, "ROUGE-SS: A New ROUGE Variant for the Evaluation of Text Summarization," *Recent Advances in Computer Science and Communications*, vol. 17, Jun. 2024, doi: 10.2174/0126662558304595240528111535.

[41] F. V. P. Samosir, H. Toba, and M. Ayub, "BESKlus : BERT Extractive Summarization with K-Means Clustering in Scientific Paper," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 1, Apr. 2022, doi: 10.28932/jutisi.v8i1.4474.