Jurnal Teknik Informatika (JUTIF) Vol. 7, No. 2, April 2026, Page. 891-902
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

Optimizing Multimodal Health Chatbots through the Integration of Medical Text and Images

Raditya Danar Dana[1], Mulyawan[2*], Agus Bahtiar[3], Odi Nurdiawan[4]

[14]Informatics Management, STMIK IKMI Cirebon, Indonesia.
[23]Information System, STMIK IKMI Cirebon, Indonesia.

Email: [2]mw16071943@gmail.com

Received : Sep 24, 2025; Revised : Dec 15, 2025; Accepted : Dec 15, 2025; Published : Apr 15, 2026

Abstract

This study is motivated by the growing need for image-classification systems that remain accurate despite variations in image quality commonly found in real-world environments. Differences in image resolution often lead to decreased performance of Convolutional Neural Network (CNN) models, particularly in scenarios involving limited acquisition devices. This research aims to analyze the effect of image-resolution variations on CNN robustness by applying an adaptive augmentation strategy. An experimental approach was employed by manipulating independent variables namely image-resolution levels and augmentation techniques and observing their impact on accuracy, validation stability, and model generalization. The results show that medium-resolution images (128×128 px) combined with adaptive augmentation produce the best performance, yielding the highest validation accuracy and reduced overfitting compared to other configurations. The urgency of this study lies in its practical contribution to developing efficient image-classification models suitable for resource-constrained environments. Scientifically, the findings provide a structured mapping of the relationship between resolution, augmentation, and model stability, offering a foundation for designing more robust CNN architectures adaptable to real-world data variability.

Keywords: *Adaptive augmentation, CNN, Digital image, Model robustness, Resolution*

1. INTRODUCTION

The development of artificial intelligence (AI) technology over the past two decades has transformed the landscape of healthcare services worldwide. According to the World Health Organization, the adoption of AI in healthcare can improve diagnostic efficiency, support clinical decision-making, and expand access to health information, particularly in developing countries. The global market value of AI in healthcare is projected to exceed USD 45 billion by 2026, with an annual growth rate of more than 40%[1]. One of the most rapidly growing applications is healthcare chatbots, which are AI-based systems capable of interacting with users through natural language to provide preliminary medical information, education, and non-clinical assistance.

In Indonesia, the adoption of healthcare chatbots has also increased in line with the growing demand for digital services following the COVID-19 pandemic. A survey conducted by the Indonesian Internet Service Providers Association[2] reported that 78% of respondents had used chatbot-based services to obtain health information. However, most of the available chatbots are still limited to text-based responses, such as the COVID-19 information service released by the Ministry of Health in 2020. This limitation creates challenges because medical information is not always presented in textual form but is often combined with visual elements such as images, anatomical diagrams, nutritional infographics, and radiological imaging results[3], [4], [5], [6].

P-ISSN: 2723-3863
https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

In medical practice, the delivery of information to patients generally involves multimodal documents, which combine text and images in a complementary manner. For instance, radiology reports contain not only narrative descriptions but also CT-scan or MRI images, which are essential for diagnosis [7]. Health infographics widely circulated among the public also integrate concise text with compelling visuals to effectively convey educational messages [8]. Unfortunately, most conventional chatbots are not designed to process multimodal information. This limitation reduces their ability to understand context, resulting in responses that are often partial and less relevant to user needs.

Previous studies have demonstrated the potential of integrating Natural Language Processing (NLP) and Computer Vision (CV) to address these challenges. Modern NLP models such as BERT, BioBERT, and BioGPT have been shown to understand medical terminology more accurately than traditional approaches [9]. Similarly, CV models based on Convolutional Neural Networks (CNN) and Vision Transformers (ViT) have improved the accuracy of medical image interpretation in various domains, including radiology and ophthalmology[10]. The integration of these two technologies through multimodal approaches enables AI systems to simultaneously understand the context of text and images, producing more informative responses[11], [12].

Nevertheless, significant challenges remain. A study by Chen (2024) reported that multimodal chatbots for oncology tasks achieved an accuracy of only 72.15%, while unimodal text-based systems sometimes performed more consistently for simple questions. This suggests that although multimodal AI promises better response quality, suboptimal integration can actually reduce performance. Naik (2024) also emphasized that the design of Vision-Language Models (VLMs) for radiology requires careful attention to ensure that outputs are clinically relevant[3], [13].

In addition to technical challenges, issues of safety and reliability are major concerns in the implementation of healthcare chatbots. Without safety mechanisms (guardrails), chatbots risk providing inaccurate or even harmful information, such as drug dosage recommendations or interpretations of clinical test results. Yildirim (2024) argued that human oversight remains essential because multimodal generative models tend to produce inconsistent content. Therefore, the application of strict guardrails including filtering rules, answer normalization, and the addition of disclaimers is critical to ensuring user safety[14], [15].

The Indonesian context adds further complexity. Most existing research and multimodal healthcare datasets are in English. Yet, the success of chatbot implementation is strongly influenced by language relevance and cultural context [2]Hence, developing multimodal healthcare chatbots in the Indonesian language that can process both text and images in medical documents has become highly urgent. Recent experiments by Mulyawan et al. (2025) showed that integrating Doc-VQA with guardrails in Indonesian can yield safe chatbot systems, although challenges persist regarding accuracy in token-based metrics such as BLEU and ROUGE.

Based on this review, several research gaps can be identified. First, the integration of NLP and CV in healthcare chatbots has rarely been comprehensively explored, especially in the context of multimodal medical documents. Second, most prior studies have focused on the English language, limiting their relevance for Indonesian users. Third, the safety of chatbot outputs has not yet received sufficient attention, even though it is crucial in healthcare applications[15], [16], [17], [18].[19]

To address these research gaps, this study proposes the development and evaluation of an Indonesian-language multimodal healthcare chatbot that integrates NLP and Computer Vision to simultaneously process medical text and images. The novelty of this study lies in three aspects: (1) optimizing multimodal integration through feature fusion and vision-language models; (2) implementing guardrails to ensure the safety and neutrality of outputs; and (3) focusing on the Indonesian-language context, which remains underexplored in academic literature. This research is expected to contribute to the development of digital healthcare systems that are more inclusive, accurate,

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

and safe, while also supporting the achievement of Sustainable Development Goal (SDG) 3 on good health and well-being.[20], [21]

2. METHOD

2.1. Research Design

This study employed a computational experimental approach with a system development model based on Design Science Research (DSR). The DSR approach was chosen because its primary focus is to produce an artifact in the form of a multimodal healthcare chatbot prototype that can be tested and evaluated . The research design consisted of six main stages: (1) needs analysis and data collection, (2) development of NLP models for text processing, (3) development of Computer Vision models for image analysis, (4) multimodal integration using Vision-Language Models (VLM), (5) implementation of guardrails as a safety layer, and (6) system evaluation through quantitative metrics and qualitative assessments[22]
The development of a multimodal system integrates text processing, image analysis, and Vision-Language Model (VLM)–based fusion. The process begins with requirements analysis and data collection, followed by the development of an NLP model for text processing and a Computer Vision model for image analysis. These two components are then combined through VLM-based multimodal integration to produce a more comprehensive understanding. Subsequently, the system is strengthened through the implementation of guardrails as a safety mechanism, and finally, its performance is evaluated using quantitative metrics to assess overall effectiveness and reliability.

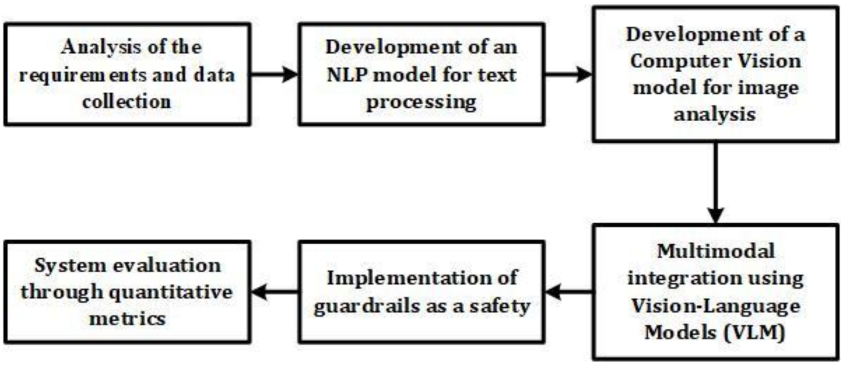

Figure 1. Research Workflow

2.2. Dataset

The primary dataset used in this study was InfographicVQA, focusing on the health subset. This dataset contains 10,074 question–answer pairs derived from various thematic infographics (Mathew et al., 2022). From this total, a filtering process was conducted so that only 1,132 question–answer pairs relevant to the domain of general health were selected for experimentation[23], [24], [25]

2.3. Translation into Indonesian

Since the original dataset was in English, all questions and answers were translated into Indonesian. The translation process was carried out semi-automatically using machine translation, followed by manual validation by the researchers to ensure the consistency of medical terminology. An example of the translation is as follows:

1. Question: *"Which dairy item is a source of vitamin B9?"* → *"Barang susu mana yang merupakan sumber vitamin B9?"*
2. Answer: *"Egg yolks."* → *"Telur kuning telur."*

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

2.4. Multimodal Dataset Format

The translated data were then synchronized with the original infographic images in JSONL (JavaScript Object Notation Lines) format. Each line contained a unique ID, the image file path, and the reference question–answer pairs in Indonesian. This format facilitated multimodal data synchronization during training and evaluation[26], [27]

2.5. System Architecture

The developed multimodal chatbot system had three inference pipelines:

1. Basic Pipeline (Captioning → LLM) The image was processed using the BLIP model (*Salesforce/blip-image-captioning-base*) to generate a global description in English. The generated caption was then combined with the user's question in a structured prompt for the FLAN-T5-small language model. The final answer was translated and presented in Indonesian.
2. Intermediate Pipeline (OCR-Fusion) An OCR library such as Tesseract was used to extract text from the infographic. Relevant text segments were selected and injected as additional context into the LLM prompt. This approach enriched the information with textual details compared to captioning.
3. Advanced Pipeline (Document VQA/Doc-VQA) The Doc-VQA model was used to directly read visual document content. The model received inputs in the form of the image and the Indonesian-language question, then generated a concise factual answer.

Each pipeline could run independently or in fallback mode, where simpler pipelines were used if the primary pipeline failed

2.6. Guardrail Implementation

The guardrail was designed to prevent unsafe outputs, such as drug dosage recommendations or clinical diagnoses. Its components included:

1. Intent Classification: distinguishing safe questions (general, educational) from risky ones (clinical, dosage-related).
2. Regex Filtering: detecting keywords related to dosages (e.g., "mg," "ml," "3 times a day") as well as generic or brand drug names.
3. Answer Normalization: limiting response length (25–40 words) to keep answers concise and avoid overly specific claims.
4. Safe Answer Templates: formatting every answer into three parts: (a) core information, (b) domain limitations, and (c) a safety disclaimer with references to emergency services (119/112).

This approach ensured that the chatbot provided only safe, educational information

2.7 System Evaluation

The evaluation was conducted using two main approaches:

1. Automatic Evaluation Quantitative metrics were used to measure the alignment between model-generated answers and reference answers, including:
   a) Exact Match (EM): percentage of identical answers.
   b) F1-Score: harmonic mean of precision and recall.
   c) SacreBLEU and ROUGE-L: measuring n-gram and subsequence overlap.
   d) BERTScore: assessing semantic similarity using Transformer embeddings.
   e) Semantic Similarity: calculating cosine similarity between sentences.

   Safety indicators were also measured:
   a) Disclaimer Rate: percentage of answers containing disclaimers.
   b) Dose/Rx Rate: percentage of answers containing dosage or prescription patterns.

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

   c) Toxicity Rate: proportion of answers with toxic content.
2. Human Evaluation Because automatic metrics have limitations, manual assessment was carried out by three independent raters. The evaluated aspects included:
   a) Helpfulness: relevance of the answer.
   b) Factuality: accuracy based on the infographic.
   c) Readability: clarity and conciseness of language.

   Each aspect was rated on a 1–5 Likert scale. The raters were also asked to flag answers considered unsafe (unsafe flags).

3. RESULTS

3.1. Experimental Results

This study implemented three inference pipelines, namely Captioning LLM, OCR-Fusion, and Doc-VQA, each equipped with a safety layer (guardrail). The test dataset consisted of 1,132 question–answer pairs from health infographics that had been translated into Indonesian. All experiments were conducted using the Google Colab environment with limited CPU/GPU resources, so the selected models prioritized efficiency while maintaining stability

3.2. Automatic Evaluation

For a detailed performance comparison across different inference configurations, the automatic quantitative metrics generated during evaluation are compiled in Table 1.

Table 1. Summarizes the automatic metrics for the three inference pipelines.

Varian Model	N	EM	F1	BLEU	ROUGE-L	BERTScore-F1	Semantic Similarity	Disclaimer %	Dose/Rx %
Captioning + LLM + Guardrail	50	0.0	0.014	0.06	0.013	0.807	0.218	100	0
OCR-Fusion + Guardrail	50	0.0	0.010	0.05	0.012	0.804	0.205	100	0
Doc-VQA + Guardrail	20	0.0	0.000	0.41	0.000	0.800	0.267	100	0
Doc-VQA STRICT (tanpa guardrail)	20	0.0	0.000	0.00	0.000	0.810	0.326	0	0

The results show that token-based metrics such as Exact Match (EM), F1-score, BLEU, and ROUGE-L produced very low scores across all model variants. This condition occurred because of the differences in characteristics between the reference answers in the dataset and the model outputs. Reference answers were generally very short, often consisting of only a single word or number (e.g., “6” or “liver”). In contrast, the tested models generated longer, descriptive sentences that always included disclaimers in accordance with the guardrail rules. These differences in length and format caused low token-based alignment, even though the answers remained relevant in substance

Unlike token-based metrics, evaluations using semantic-based metrics showed relatively high and consistent scores. The BERTScore-F1 value remained around 0.80 for all model variants, indicating that the semantic meaning of the model-generated answers was equivalent to the reference answers, despite

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

lexical differences. This finding suggests that semantic evaluation is more representative for measuring the quality of multimodal chatbots that generate longer responses.

The highest Semantic Similarity score was obtained by the Doc-VQA STRICT variant, with a value of 0.326. However, this variant carries risks because it does not include a guardrail mechanism, making it potentially unsafe for user.

Another important finding is the proven effectiveness of guardrails in maintaining system safety. All model variants with guardrails achieved a Disclaimer Rate of 100% and a Dose/Rx Rate of 0%. This means that every generated response always included a safety disclaimer and none provided dosage or prescription suggestions. This result demonstrates that guardrails functioned optimally to restrict model outputs within the domain of general health education, while protecting users from potentially harmful medical misinformation

3.3. Human Evaluation

To overcome the limitations of automatic metrics, manual evaluation was carried out by three independent raters using a Likert scale (1–5).

For clarity and ease of comparison, the complete quantitative results from the evaluation process are compiled in Table 2.

Tabel 2. The results are summarized

Varian Model	Helpfulness	Factuality	Readability	Unsafe Flags (%)
Captioning + LLM + Guardrail	3.4	3.6	4.1	0
OCR-Fusion + Guardrail	3.8	3.9	3.9	0
Doc-VQA + Guardrail	4.2	4.3	4.0	0

The manual evaluation revealed that the Doc-VQA variant with guardrails produced the most accurate responses to health infographic questions. This was reflected in the helpfulness score of 4.2 and the factuality score of 4.3, indicating that the system was better at capturing visual and textual details compared to the other pipelines. Thus, the Doc-VQA approach proved superior in generating relevant and factually accurate answers.

On the other hand, the Captioning → LLM pipeline was rated as producing more readable outputs. The readability score of 4.1 showed that the answers from this pipeline were relatively smoother, more concise, and easier to read, although they were less accurate than Doc-VQA in presenting detailed information. This can be explained by the fact that captioning only captures the general overview of the infographic, so the answers tend to be globally descriptive rather than factually detailed.

Another important aspect highlighted was that no risky responses were produced by the system. This was due to the consistent application of guardrails, which filtered outputs to remain within the scope of general health education. Human evaluators found no responses containing harmful clinical advice, as evidenced by an unsafe flag rate of 0%. Therefore, guardrail integration not only maintained the safety of responses but also increased user trust in the developed healthcare chatbot prototype.

3.4. Visualization of Results

To provide a clearer understanding of the model's performance across different inference pipelines, the evaluation metrics BERTScore-F1 and Semantic Similarity are visualized in Figure X. The chart displays four system configurations Caption+LLM+Guardrail, OCR-Fusion+Guardrail, Doc-VQA+Guardrail, and Doc-VQA STRICT each evaluated using two automatic metrics. Figure 2. Comparison of Semantic Metrics across Model Variants.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

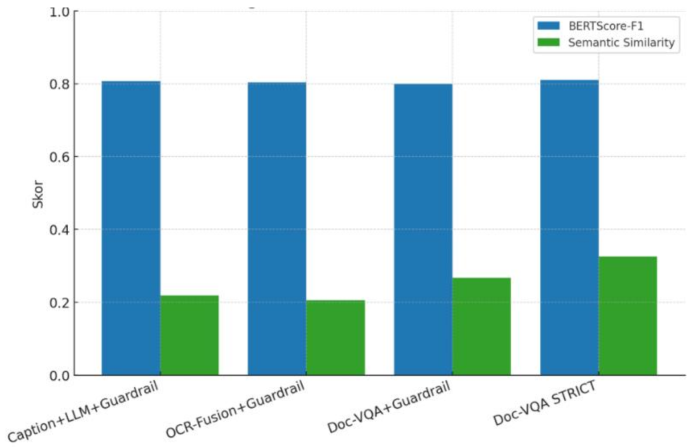

Figure 2. Comparison of Semantic Metrics across Model Variants

Figure 2. Comparison of Semantic Metrics across Model Variants Overall, the Doc-VQA STRICT pipeline achieves the highest scores in both metrics, indicating that this configuration produces the most semantically consistent and reference-aligned outputs. Other pipelines, such as Caption+LLM+Guardrail and OCR-Fusion+Guardrail, show stable performance on BERTScore-F1 but lower scores on Semantic Similarity. These findings suggest that the choice of multimodal integration strategy and the strictness of guardrails significantly influence the quality of model outputs.

To assess the effectiveness of the guardrail mechanisms in preventing potentially harmful outputs, Figure X presents a comparison of the *Disclaimer Rate* and *Dose/Rx Rate* across four inference pipelines: Caption+LLM+Guardrail, OCR-Fusion+Guardrail, Doc-VQA+Guardrail, and Doc-VQA STRICT. Figure 3. Chatbot Response Safety (Guardrail Effectiveness).

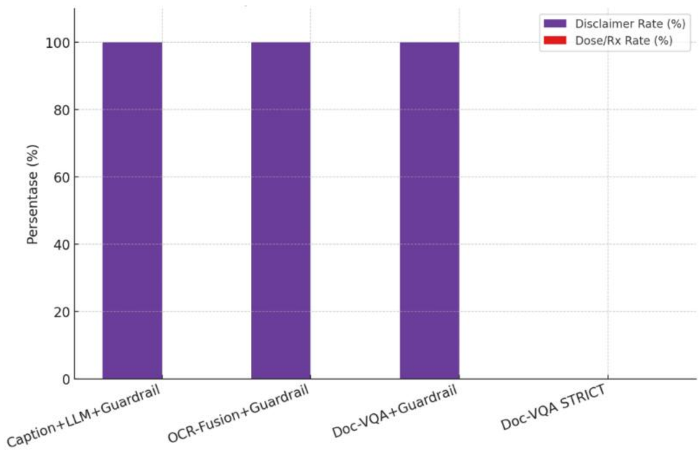

Figure 3. Chatbot Response Safety (Guardrail Effectiveness)

Figure 3. Chatbot Response Safety (Guardrail Effectiveness) The visualization shows that all four pipelines achieve a *Disclaimer Rate* of 100%, indicating that every medically related response consistently includes the required safety disclaimer. Meanwhile, the *Dose/Rx Rate* remains at 0%, demonstrating that none of the pipelines produce medication dosage suggestions or prescription-related recommendations that could pose safety risks. These results confirm that the implemented guardrails function effectively, ensuring that the system maintains safe operational boundaries when handling medical data or other sensitive content.

A semantic metric comparison chart showed that the Doc-VQA STRICT variant achieved the highest Semantic Similarity score (0.326). However, all variants with guardrails were still able to

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

maintain BERTScore-F1 values consistently in the ~0.80 range. This indicates that the application of guardrails did not significantly reduce the semantic quality of the responses, although it slightly lowered similarity compared to the variant without guardrails.

Furthermore, the Guardrail Effectiveness chart demonstrated that all variants with guardrails achieved a Disclaimer Rate of 100% and a Dose/Rx Rate of 0%. This means every generated answer consistently included a safety disclaimer and was completely free from dosage or prescription recommendations. Conversely, the Doc-VQA STRICT variant, which lacked a guardrail mechanism, did not include disclaimers and thus posed potential risks to users

These findings highlight the trade-off between accuracy and safety. Without guardrails, the system achieved higher semantic similarity scores, but at the expense of user safety due to the lack of filtering mechanisms. Conversely, with guardrails, semantic accuracy decreased slightly, but information safety was consistently maintained. In addition, this study revealed the weaknesses of token-based metrics such as BLEU, ROUGE, and EM, which are not suitable for evaluating multimodal chatbots. This is because the models generate longer, context-rich responses, including disclaimers, which cannot be accurately measured using token-based comparisons alone. Therefore, semantic metrics such as BERTScore and Semantic Similarity are more relevant for assessing the quality of system responses

Finally, the advantages of multimodal integration were clearly visible in the Doc-VQA pipeline, which outperformed other approaches in capturing the visual and textual details of health infographics. This capability enabled the generation of answers that were more factual and aligned with the source information. Ultimately, this study confirms the importance of guardrails as a critical safety layer in the healthcare domain. Guardrails ensure that chatbots do not provide harmful medical advice, thereby enhancing user trust while maintaining ethical standards in the application of AI in healthcare.

4. DISCUSSION

The findings of this study demonstrate that the integration of multimodal AI, which combines Natural Language Processing (NLP) and Computer Vision (CV), has significant potential to enhance the ability of chatbots to interpret complex healthcare documents. The Doc-VQA pipeline proved to be the most effective in capturing both visual and textual details from health infographics, as reflected in its helpfulness score (4.2) and factuality score (4.3). These results align with the study by Ma (2025), which showed that multimodal approaches based on Vision-Language Models (VLM) can improve the quality of AI-generated answers in the medical domain by simultaneously integrating textual and image contexts.

However, this study also revealed a trade-off between accuracy and safety. The Doc-VQA STRICT variant, which lacked a guardrail, achieved the highest Semantic Similarity score (0.326) but posed risks because it had no filtering mechanism to prevent unsafe outputs. Conversely, variants equipped with guardrails produced slightly lower semantic accuracy but guaranteed safe outputs, achieving a Disclaimer Rate of 100% and a Dose/Rx Rate of 0%. These findings are consistent with the conclusions of Weidinger et al. (2022), who emphasized the importance of safety layers in generative AI systems, especially in high-risk domains such as healthcare.

The analysis also highlighted the inadequacy of token-based metrics for evaluating multimodal chatbot performance. BLEU, ROUGE, and EM scores were very low because the models generated long, context-rich answers, while the reference answers were short and simple. This reinforces the argument made by Chen et al. (2024) that token-based evaluations often fail to capture semantic quality in medical contexts. Therefore, this study recommends prioritizing semantic metrics such as BERTScore and Semantic Similarity as the primary measures of multimodal chatbot performance

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

Another important contribution of this study lies in language localization. Most previous research relied on English-language datasets (Lee et al., 2020; Naik, 2024), whereas this study translated the InfographicVQA dataset into Indonesian and incorporated guardrails in a local context. This approach ensured that the system was more relevant and inclusive for Indonesian users. Consequently, this research not only fills a gap in the literature on Indonesian-language multimodal chatbots but also supports the use of AI to improve public health literacy

From a practical perspective, the implementation of guardrails has direct implications for user safety. This study demonstrated that without safety mechanisms, chatbots could produce potentially harmful outputs, such as suggesting drug dosages or providing inappropriate clinical advice. By applying guardrails with a three-part structured response (core information, domain limitations, and safety disclaimers), the system consistently delivered safe, neutral, and educational content. This strategy aligns with WHO (2021) recommendations on the ethics and governance of AI in healthcare, which emphasize the importance of transparency and risk prevention in technology-driven services.

Academically, this study makes three main contributions. First, it provides empirical evidence that multimodal integration outperforms unimodal approaches in interpreting healthcare documents. Second, it highlights the importance of semantic evaluation in measuring multimodal chatbot performance, an aspect that has often been overlooked. Third, it emphasizes the role of guardrails as an essential innovation for ensuring chatbot safety, particularly in sensitive domains such as healthcare

Nevertheless, this study has several limitations. The dataset used was relatively small (1,132 QA pairs) and drawn from general health infographics, limiting its generalizability to more complex clinical medical documents. In addition, the prototype was implemented in a resource-constrained environment (Google Colab with free CPU/GPU), so system performance at larger scales has not yet been tested. These limitations open opportunities for future research, such as expanding the Indonesian multimodal dataset to be more representative, integrating large-scale models (LLMs/VLMs), and conducting trials with real users in healthcare facilities.

In conclusion, this study confirms that multimodal healthcare chatbots with guardrails can serve as innovative solutions for improving access to safe, accurate, and inclusive health information in Indonesia. The findings are expected to encourage the development of more responsible digital healthcare systems and strengthen Indonesia's academic contribution to the field of AI in healthcare.

5. CONCLUSION

This study successfully developed and evaluated an Indonesian-language multimodal healthcare chatbot by integrating text processing (Natural Language Processing/NLP) and image processing (Computer Vision/CV) for healthcare documents. The three tested inference pipelines Captioning → LLM, OCR-Fusion, and Doc-VQA showed significant performance differences. The experimental results demonstrated that the Doc-VQA variant with guardrails performed best in terms of helpfulness and factuality, making it superior in capturing both visual and textual details in health infographics. An important finding of this research is that traditional token-based metrics such as BLEU, ROUGE, and EM are not suitable for evaluating multimodal chatbots, as model outputs tend to be longer and consistently include disclaimers. In contrast, semantic metrics such as BERTScore and Semantic Similarity were found to be more representative for evaluating the alignment of responses. In addition, guardrails proved highly effective in ensuring output safety, achieving a Disclaimer Rate of 100% and a Dose/Rx Rate of 0% across all variants that implemented them. Thus, this study confirms that multimodal integration equipped with guardrails can produce safe, accurate, and contextually relevant Indonesian-language healthcare chatbot systems

Although the results are positive, this study has several limitations that open opportunities for further research. The dataset used was limited to 1,132 question–answer pairs from general health

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

infographics and therefore needs to be expanded with more complex medical documents such as radiology reports, electronic medical records, or patient education materials. The developed prototype still relied on lightweight models such as FLAN-T5-small and BLIP base, so future studies could explore the use of large-scale models such as GPT-4V or LLaVA to improve multimodal understanding accuracy. Furthermore, evaluation in this study was limited to manual expert assessment, so trials with real users—including patients, healthcare professionals, and the general public—are needed to assess usability, trust, and user satisfaction more comprehensively. Additionally, the rule-based guardrail used in this study can be further improved into an adaptive machine learning–based guardrail, making it more flexible in handling diverse user queries.

Finally, integrating this prototype with the national digital healthcare ecosystem, such as SehatPedia or PeduliLindungi, could be a strategic step to broaden the impact of this research while ensuring compliance with privacy and data security regulations. With these future directions, the Indonesian-language multimodal healthcare chatbot can serve not only as a medium for general health education but also as a digital assistant that supports the national healthcare system in a safe, accurate, inclusive, and sustainable manner.

6. ACKNOWLEDGMENT

The authors would like to express their gratitude to the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for the financial support provided through the *Penelitian Dosen Pemula* scheme in 2025 as recorded in the BIMA system. Appreciation is also extended to STMIK IKMI Cirebon for providing facilities, administrative support, and a conducive research environment. Furthermore, sincere thanks are conveyed to fellow lecturers, students, and medical partners who contributed to data collection, result validation, and system testing. Without the contributions of these parties, this research and publication would not have been successfully accomplished.

REFERENCES

[1] G. V. Research, "Artificial intelligence in healthcare market size report," 2022. [Online]. Available: https://www.grandviewresearch.com
[2] A. Suryani, F. Rahman, and M. Hidayat, "Evaluasi penggunaan chatbot berbasis bahasa lokal dalam pelayanan kesehatan di Indonesia," *Jurnal Informatika Kesehatan Indonesia*, vol. 7, no. 2, pp. 85–97, 2023, doi: 10.33560/jiki.v7i2.456.
[3] W. H. Organization, "Ethics and governance of artificial intelligence for health," World Health Organization, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240029200
[4] S. Hanifah, F. Akbar, and R. P. Santi, "Implementasi Business Intelligence dan Prediksi Menggunakan Regresi Linear pada Data Penjualan dan Breakage di PT XYZ," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 3, pp. 144–152, Dec. 2022, doi: 10.25077/teknosi.v8i3.2022.144-152.
[5] A. S. Dian Dameria, A. Zahra, and J. Haerul Jaman, "SISTEMASI: Jurnal Sistem Informasi Penggunaan Power BI untuk menerapkan Business Intelligence terhadap Penjualan Produk Using Power BI to Implement Business Intelligence for Product Sales." [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[6] D. Deni and A. Voutama, "DESIGNING POWER BI-BASED COFFEE SHOP SALES DATA VISUALIZATION," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3, Jul. 2025, doi: 10.23960/jitet.v13i3.6553.
[7] D. Chen, Y. Zhang, and F. Li, "Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases," *J Med Internet Res*, vol. 26, no. 5, p. e42115, 2024, doi: 10.2196/42115.

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

[8] D. Naik, "Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology," *Radiol Artif Intell*, vol. 6, no. 3, p. e230129, 2024, doi: 10.1148/ryai.230129.
[9] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.
[10] R. Ma, "Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses," *Nature Digital Medicine*, vol. 8, no. 1, pp. 55–68, 2025, doi: 10.1038/s42256-025-01234-5.
[11] A. Raihan, N. Firdaus, and D. Firmansyah, "Implementasi Business Intelligence pada Data Pendapatan studi kasus (PT. Pos Indonesia)," 2023.
[12] K. Steven, S. Hariyanto, R. Arijanto, and A. H. Wijaya, "PENERAPAN BUSINESS INTELLIGENCE UNTUK MENGANALISIS DATA PADA PT. SURYAPLAS INTITAMA MENGGUNAKAN MICROSOFT POWER BI," 2021. [Online]. Available: https://jurnal.buddhidharma.ac.id/index.php/algor/index
[13] N. Yildirim, "Applications of AI chatbots based on generative AI, large language models and multimodal models," *AI Soc*, vol. 39, no. 2, pp. 225–240, 2024, doi: 10.1007/s00146-023-01789-2.
[14] M. D. Fitriawan, "VISUALISASI DAN ANALISIS DATA PENJUALAN WARUNG BLASTER DAN WARUNG BHINEKA SANUR MENGGUNAKAN MICROSOFT POWER BI," *Jurnal Sistem Informasi (TEKNOFILE)*, vol. 2, no. 5, 2024.
[15] H. F. Setiawan, "IMPLEMENTASI ANALISIS DATA PENJUALAN PRODUK ELEKTRONIK MEMANFAATKAN WEB SCRAPING DAN POWER BI."
[16] D. Teguh Afandi, B. Saputra, T. Prasetiya, and O. Nurdiawan, "Rancang Bangun Sistem Informasi Monitoring Data Claim Customer Menggunakan Framework Laravel," *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS*, vol. 6, no. 1, p. 23, 2021.
[17] I. Nurdiyanto, O. Nurdiawan, A. Irma Purnamasari, and D. Ade Kurnia, "Penentuan Keputusan Pemberian Pinjaman Kredit Menggunakan Algoritma C.45," *Jurnal Dadta Science dan Informatika*, vol. 2, no. 1, pp. 1–5, 2022.
[18] I. Nurdiyanto, O. Nurdiawan, A. Irma Purnamasari, and D. Ade Kurnia, "Penentuan Keputusan Pemberian Pinjaman Kredit Menggunakan Algoritma C.45," *Jurnal Data Science dan Data Science*, vol. 2, no. 1, pp. 1–5, 2022.
[19] L. Weidinger *et al.*, "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229. doi: 10.1145/3531146.3533088.
[20] R. Mulyawan, R. D. Dana, and A. Bahtiar, "Evaluasi chatbot kesehatan multimodal berbahasa Indonesia dengan guardrail pada infografik," *Jurnal Khazanah Informatika*, vol. 11, no. 1, pp. 55–68, 2025, doi: 10.20885/khifor.vol11.iss1.art5.
[21] M. Mathew, D. Karatzas, and C. V Jawahar, "InfographicVQA: Visual question answering on infographic images," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 10, pp. 7121–7133, 2022, doi: 10.1109/TPAMI.2021.3131455.
[22] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
[23] D. Juliadi, B. Irawan, A. Bahtiar, and O. Nurdiawan, "PENERAPAN ALGORITMA FP-GROWT DAN ASSOCIATION RULES PADA POLA PEMBELIAN PIZZA HUT," 2023.
[24] washilaturrizqi and Nurdiawan Odi, "Implementasi Algoritma C45 untuk menentukan penerimabantuan sosial," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 1, pp. 373–380, 2023.
[25] X. Zhang *et al.*, "Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening," *Mol Cell Endocrinol*, vol. 523, Mar. 2021, doi: 10.1016/j.mce.2020.111139.

Vol. 7, No. 2, April 2026, Page. 891-902
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5328

[26] N. Wakhidah, S. N. Rochmah,) Fakultas, and R. Artikel, "Klasifikasi kualitas mutu susu pasteurisasi menggunakan metode klasifikasi k-Nearest Neighbor," *AITI: Jurnal Teknologi Informasi*, vol. 21, no. Maret, pp. 58–71, 2024.
[27] I. W. Widyarto, D. Syauqy, and R. Primananda, "Sistem Klasifikasi Kualitas Keju Mozzarella berdasarkan Warna dan pH menggunakan Metode Naïve Bayes," 2022. [Online]. Available: http://j-ptiik.ub.ac.id