ABSTRACT

PT Bank Rakyat Indonesia (BRI) faces significant challenges in processing and accessing information from large volumes of unstructured multimodal documents containing both text and images. Manual systems or traditional Retrieval Augmented Generation (RAG) approaches are often inefficient and can lose crucial visual context. This research aims to develop a Multimodal RAG (MRAG) *system to enhance the accuracy and efficiency of information extraction from such documents at BRI. The research is conducted using the CRISP-DM framework (Cross-Industry Standard Process for Data Mining), which includes business and data understanding, data preparation, modeling, evaluation, and system deployment. The proposed system adapts the ColPali approach, implemented as ColQwen2.5, for document retrieval by treating each document page as an image. Furthermore, it utilizes a fine-tuned Vision Language Model (VLM), Qwen2.5-VL, for contextually relevant answer generation. A key contribution of this study is the creation of a new Indonesian image-question-answer dataset, collected from visual public documents, comprising 1.318 unique images and 3.930 question-answer pairs. Evaluation results demonstrate significant performance: the ColQwen2.5 retriever model achieved an MRR@5 of 0,92762, while the fine-tuned Qwen2.5-VL generator model attained a BERT-F1 score of 0,8534 and an LLM-Eval (using GPT-4o) accuracy of 0,8603, marking a 3.3% improvement over its base model. The development of this* MRAG *system offers considerable potential for optimizing knowledge management and supporting improved decision-making processes at PT Bank Rakyat Indonesia.*

Keywords— ***Multimodal Retrieval Augmented Generation, ColPali, Vision Language Model, Information Retrieval, Visual Document Understanding.***