UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

ANALISIS KOMPARASI ARSITEKTUR *RETRIEVAL-AUGMENTED GENERATION* (RAG) BERBASIS MODEL LLAMA 3.1 PADA IMPLEMENTASI SISTEM *CHATBOT* PEMBANTU PENGAMBILAN KEPUTUSAN INVESTASI

Ryvalino Dhanu Ekaputra
22/499364/SV/21310

Dengan meningkatnya indeks pasar saham global beberapa tahun ke belakang, serta kemudahan *investor* dalam mengakses pasar saham, diperlukan kemampuan untuk mengambil keputusan investasi yang tepat. Integrasi teknologi kecerdasan buatan dapat membantu proses pengambilan keputusan investasi tersebut. Penelitian ini mengembangkan sistem *chatbot* berbasis *large language model* (LLM) dengan model *generator* llama3.1:8b. Sistem akan menggunakan data dokumen finansial yang tersedia pada *database* EDGAR, serta menggunakan *dataset* financeBench untuk *benchmarking*. Evaluasi dilakukan pada arsitektur *hybridRAG* dan *vectorRAG* dengan variasi metode segmentasi teks dan *embedding model* yang digunakan, yaitu BGE-M3 dan octen-embedding-0.6B. Berdasarkan hasil evaluasi, arsitektur *vectorRAG* dengan metode segmentasi teks berbasis *token* serta *embedding model* BGE-M3 merupakan konfigurasi sistem dengan hasil paling optimal, dengan skor s*emantic similarity* sebesar 0,6147, skor BERT *Recall* sebesar 0,7889, skor BERT *Precision* sebesar 0,7161, skor BERT F1 sebesar 0,7481, serta latensi 12,52 detik. Hasil ini mengungguli performa arsitektur *hybridRAG* dengan skor *semantic similarity* sebesar 0,6045, skor BERT *Recall* 0,7891, skor BERT *Precision* sebesar 0,7056, skor BERT F1 0,7424, dan latensi 6,48 detik. Sistem kemudian diimplementasikan menggunakan React.Js, Vite, FastAPI, ChromaDB dan Neo4J. Hasil implementasi menunjukkan keberhasilan pada 17 skenario *blackbox testing* serta skor 91,2% pada *user acceptance testing* (UAT). Penelitian dan pengembangan sistem ini diharapkan dapat membantu *investor* dalam

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

memahami data krusial perusahaan dan mengambil keputusan investasi dengan lebih akurat.

Kata Kunci : *Chatbot*, Arsitektur RAG, *VectorRAG*, *HybridRAG*, ChromaDB, Neo4J, Sistem Pendukung Keputusan Investasi, Aplikasi Web

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

COMPARATIVE ANALYSIS OF RETRIEVAL-AUGMENTED GENERATION (RAG) ARCHITECTURE BASED ON THE LLAMA 3.1 MODEL IN CHATBOT IMPLEMENTATION FOR AN INVESTMENT DECISION SUPPORT SYSTEM

Ryvalino Dhanu Ekaputra
22/499364/SV/21310

With the increasing global stock market indices in recent years, as well as the ease of investor access to the stock market, the ability to make sound investment decisions is required. The integration of artificial intelligence technology can assist in the investment decision-making process. This research develops a chatbot system based on a large language model (LLM) with llama3.1:8b as the generator model. The system will utilize financial document data available in the EDGAR database, as well as the FinanceBench dataset for benchmarking. Evaluation is conducted on hybridRAG and vectorRAG architectures with variations in text segmentation methods and embedding models used, namely BGE-M3 and octen-embedding-0.6b. Based on the evaluation results, the vectorRAG architecture with token-based text segmentation method and BGE-M3 embedding model represents the most optimal system configuration, with a semantic similarity score of 0.6147, a BERT Recall score of 0.7889, a BERT Precision score of 0.7161, a BERT F1 score of 0.7481, and a latency of 12.52 seconds. These results outperform the performance of the hybridRAG architecture with a semantic similarity score of 0.6045, a BERT Recall score of 0.7891, a BERT Precision score of 0.7056, a BERT F1 score of 0.7424, and a latency of 6.48 seconds. The system is then implemented using React.Js, Vite, FastAPI, ChromaDB and Neo4J. The implementation results demonstrate success across 17 black-box testing scenarios as well as a score of 91.2% in user acceptance testing (UAT). This research and system development are expected to assist investors in understanding crucial company data and making investment decisions more accurately.

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Keywords *: Chatbot, RAG Architecture, VectorRAG, HybridRAG, ChromaDB, Neo4J, Investment Decision Support System, Web Application*