UNIVERSITAS
GADJAH MADA
IMPLEMENTASI SISTEM CHATBOT DOKUMENTASI BERBASIS RETRIEVAL-AUGMENTED GENERATION (RAG) PADA INFRASTRUKTUR CLOUD DENGAN OTOMASI SINKRONISASI KNOWLEDGE BASE
Optra Dananjaya, Ir. Budi Bayu Murti, S.T., M.T., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

IMPLEMENTASI SISTEM *CHATBOT* DOKUMENTASI BERBASIS *RETRIEVAL-AUGMENTED GENERATION (RAG)* PADA INFRASTRUKTUR *CLOUD* DENGAN OTOMASI SINKRONISASI *KNOWLEDGE BASE*

Optra Dananjaya

22/499522/SV/21339

Dokumentasi teknis yang luas dan terus diperbarui menuntut mekanisme akses informasi yang cepat, relevan, dan bersumber dari referensi resmi, sehingga pendekatan konvensional berbasis pencarian manual tidak lagi memadai. Proyek akhir ini mengimplementasikan sistem *chatbot* dokumentasi berbasis *Retrieval-Augmented Generation* (RAG) dengan studi kasus dokumentasi resmi platform Dify, yang dibangun di atas infrastruktur *cloud* menggunakan *VPS, Docker Compose*, *NGINX*, dan *HTTPS/SSL*, serta memanfaatkan Dify sebagai platform terpadu untuk pengelolaan *workflow*, *knowledge base*, *retrieval*, *reranking*, dan generasi jawaban. Akses terhadap model bahasa dilakukan melalui OpenAI dengan *text-embedding-3-small* sebagai model *embedding* dan *jina-reranker-v2-base-multilingual* sebagai model *reranking*, sedangkan pembaruan *knowledge base* dikelola secara otomatis melalui *pipeline* sinkronisasi berbasis *ETL/CDC* yang dijadwalkan menggunakan *cron job*. Proses *indexing* dilakukan dengan memecah dokumen menjadi *chunk-chunk* teks yang kemudian disimpan dalam *vector store*, sehingga pencarian konteks relevan dapat dilakukan secara efisien pada saat pengguna mengajukan pertanyaan. Hasil pengujian menunjukkan bahwa seluruh 20 skenario fungsional berhasil dijalankan, sistem memperoleh *Success Rate* 97,50% dan Indeks Kualitas Jawaban 98,75% dari 40 pertanyaan uji, serta *Retrieval Relevance Rate* 97,14% dan *MRR* 0,869 dari 35 pertanyaan berbasis *retrieval*, dengan *Consistency Rate* 100,00% dan rata-rata waktu respons 17,383 detik. *Pipeline* sinkronisasi mampu menangani operasi *insert*, *update*, dan *skip* secara otomatis tanpa *error*, sehingga secara keseluruhan sistem terbukti mampu mendukung akses informasi dokumentasi secara relevan, terstruktur, dan berbasis sumber resmi.

Kata kunci: *chatbot*, *Retrieval-Augmented Generation*, *cloud deployment*, Dify, *knowledge base*

UNIVERSITAS
GADJAH MADA
IMPLEMENTASI SISTEM CHATBOT DOKUMENTASI BERBASIS RETRIEVAL-AUGMENTED GENERATION (RAG) PADA INFRASTRUKTUR CLOUD DENGAN OTOMASI SINKRONISASI KNOWLEDGE BASE
Optra Dananjaya, Ir. Budi Bayu Murti, S.T., M.T., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

IMPLEMENTATION OF A DOCUMENTATION CHATBOT SYSTEM BASED ON RETRIEVAL-AUGMENTED GENERATION (RAG) ON CLOUD INFRASTRUCTURE WITH AUTOMATED KNOWLEDGE BASE SYNCHRONIZATION

Optra Dananjaya
22/499522/SV/21339

Extensive and frequently updated technical documentation demands an information access mechanism that is fast, relevant, and grounded in official sources, making conventional manual search approaches inadequate. This final project implements a documentation chatbot system based on Retrieval-Augmented Generation (RAG) using the official Dify documentation as the case study, deployed on cloud infrastructure comprising a VPS, Docker Compose, NGINX, and HTTPS/SSL, with Dify serving as the unified platform for managing the workflow, knowledge base, retrieval, reranking, and answer generation processes. Language model access is provided through OpenAI using text-embedding-3-small as the embedding model and jina-reranker-v2-base-multilingual as the reranking model, while knowledge base updates are managed automatically through an ETL/CDC-based synchronization pipeline scheduled via cron job. The indexing process involves splitting documents into text chunks that are stored in a vector store, enabling efficient retrieval of relevant context whenever a user submits a query. Testing results demonstrate that all 20 functional scenarios were completed successfully, with the system achieving a Success Rate of 97.50% and an Answer Quality Index of 98.75% across 40 test questions, as well as a Retrieval Relevance Rate of 97.14% and an MRR of 0.869 across 35 retrieval-based questions, alongside a Consistency Rate of 100.00% and an overall average response time of 17.383 seconds. The synchronization pipeline successfully handled insert, update, and skip operations automatically without errors, demonstrating that the system is capable of supporting documentation information access in a relevant, structured, and source-based manner.

Keywords: *chatbot, Retrieval-Augmented Generation, cloud deployment, Dify, knowledge base*