

**IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION
UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE
MODEL BERBASIS VECTOR DATABASE**



SKRIPSI

Atras Shalhan
00000050597

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2024**

**IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION
UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE
MODEL BERBASIS VECTOR DATABASE**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

Atras Shalhan

00000050597

UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2024

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Atras Shalhan

NIM : 00000050597

Program Studi : Informatika

Menyatakan dengan sesungguhnya bahwa Skripsi saya yang berjudul:
**IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION
UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE
MODEL BERBASIS VECTOR DATABASE**

merupakan hasil karya saya sendiri, bukan merupakan hasil plagiat, dan tidak pula dituliskan oleh orang lain; Semua sumber, baik yang dikutip maupun dirujuk, telah saya cantumkan dan nyatakan dengan benar pada bagian Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan skripsi maupun dalam penulisan laporan karya ilmiah, saya bersedia menerima konsekuensi untuk dinyatakan TIDAK LULUS. Saya juga bersedia menanggung segala konsekuensi hukum yang berkaitan dengan tindak plagiarisme ini sebagai kesalahan saya pribadi dan bukan tanggung jawab Universitas Multimedia Nusantara.

Tangerang, 8 Mei 2024

UNIVERSITAS
MULTIMEDIA
NUSANTARA



(Atras Shalhan)

HALAMAN PENGESAHAN

Skripsi dengan judul

IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE MODEL BERBASIS VECTOR DATABASE

oleh

Nama : Atras Shalhan
NIM : 00000050597
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

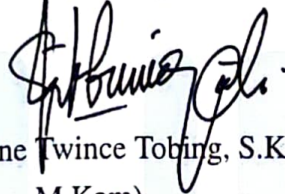
Telah diujikan pada hari Jumat, 07 Juni 2024

Pukul 10.00 s/s 12.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang



(Fenina Adline Twince Tobing, S.Kom.,
M.Kom)

NIDN: 406058802

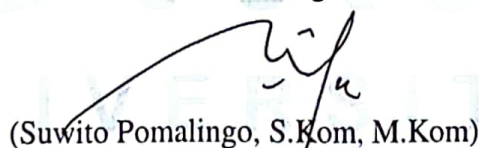
Penguji



(David Agustriawan, S.Kom., M.Sc.,
Ph.D.)

NIDN: 0525088601

Pembimbing



(Suwito Pomalingo, S.Kom, M.Kom)

NIDN: 0309008503

Pjs. Ketua Program Studi Informatika,



(Dr. Eng. Niki Prastomo, S.T., M.Sc.)

NIDN: 0419128203

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Atras Shalhan

NIM : 00000050597

Program Studi : Informatika

Jenjang : S1

Jenis Karya : Skripsi

Menyatakan dengan sesungguhnya bahwa:

- Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya di repositori Knowledge Center, sehingga dapat diakses oleh Civitas Akademika/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial dan saya juga tidak akan mencabut kembali izin yang telah saya berikan dengan alasan apapun.
- Saya tidak bersedia karena dalam proses pengajuan untuk diterbitkan ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*)**.

Tangerang, 8 Mei 2024

Yang menyatakan

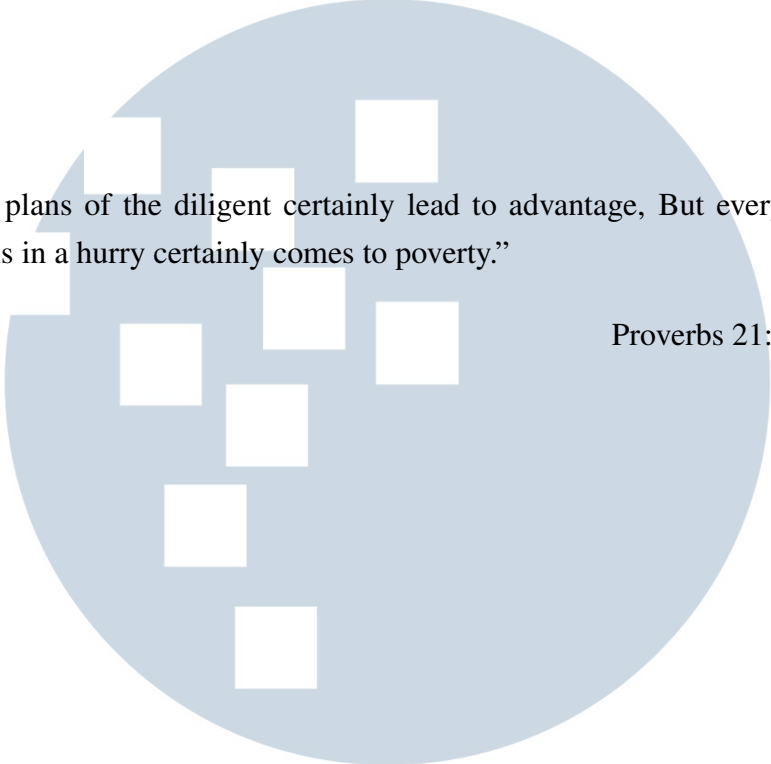


Atras Shalhan

U M M N
UNIVERSITAS
MULTIMEDIA
NUSANTARA

** Jika tidak bisa membuktikan LoA jurnal/HKI selama enam bulan ke depan, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN.

Halaman Persembahan / Motto



”The plans of the diligent certainly lead to advantage, But everyone who is in a hurry certainly comes to poverty.”

Proverbs 21:5 (NASB)

UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE MODEL BERBASIS VECTOR DATABASE dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Pjs. Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Bapak Suwito Pomalingo, S.Kom, M.Kom, sebagai Pembimbing pertama yang telah banyak meluangkan waktu untuk memberikan bimbingan, arahan dan motivasi atas terselesainya tesis ini.
5. Orang Tua dan keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tesis ini.

Semoga skripsi ini bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 8 Mei 2024



Atras Shalhan

**IMPLEMENTASI TEKNIK RETRIEVAL AUGMENTED GENERATION
UNTUK MENGHILANGKAN HALUSINASI PADA LARGE LANGUAGE
MODEL BERBASIS VECTOR DATABASE**

Atras Shalhan

ABSTRAK

Penelitian ini bertujuan untuk mengatasi masalah halusinasi pada *LLM* berbasis vektor dengan menerapkan teknik *Retrieval Augmented Generation (RAG)*. *RAG* mengintegrasikan proses pengambilan informasi dengan proses *text generation* dari *LLM* untuk meningkatkan kualitas hasil generasi dan mengurangi kemungkinan menghasilkan informasi yang tidak akurat atau bersifat halusinatif. Penelitian ini berfokus pada implementasi *RAG* pada *Large Language Model (LLM)* dengan menggunakan *vector database*. Metode ini diuji dengan mengukur akurasi dari sistem yang dihasilkan. Hasil penelitian menunjukkan bahwa implementasi *RAG* berhasil menghasilkan sistem dengan tingkat akurasi sebesar 86.84%. Penelitian ini menunjukkan potensi besar dari pendekatan *RAG* untuk mengatasi masalah halusinasi dalam *LLM* dan meningkatkan kualitas *text generation* pada *LLM* secara signifikan.

Kata kunci: *Artificial intelligence, Chatbot, Large Language Models, Retrieval Augmented Generation, Vector Database*



**IMPLEMENTATION OF AUGMENTED GENERATION RETRIEVAL
TECHNIQUE TO ELIMINATE HALLUCINATIONS IN LARGE
LANGUAGE MODEL BASED ON VECTOR DATABASE**

Atras Shalhan

ABSTRACT

This study aims to address hallucination issues in vector-based Large Language Models (LLMs) by applying Retrieval Augmented Generation (RAG) technique. RAG integrates the information retrieval process with the text generation process of LLM to enhance the quality of generated outputs and reduce the likelihood of producing inaccurate or hallucinatory information. The research focuses on implementing RAG on LLM using a vector database. This method is evaluated by measuring accuracy of the generated system. The results indicate that the implementation of RAG successfully produces a system with an accuracy rate of 86.84%. This research demonstrates the significant potential of the RAG approach in addressing hallucination issues in LLMs and significantly improving text generation quality in LLMs.

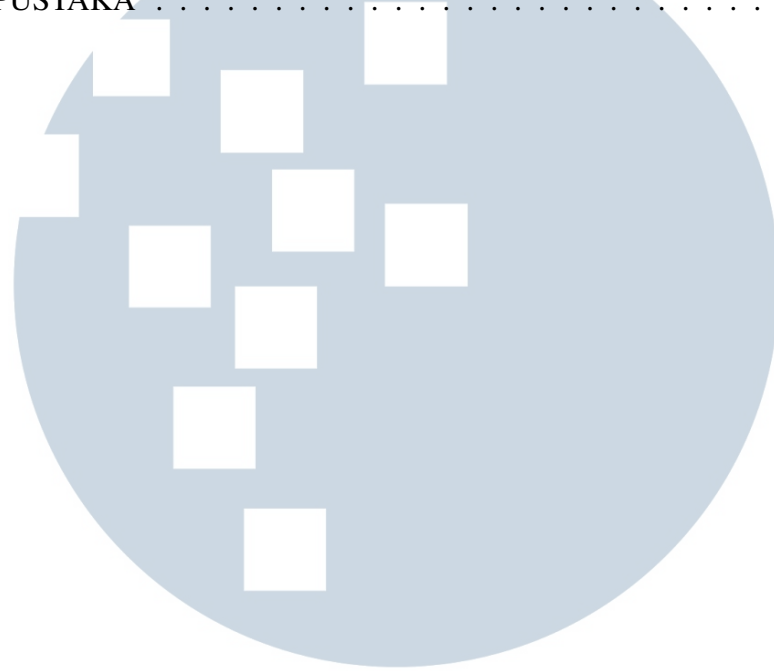
Keywords: *Artificial intelligence, Chatbot, Large Language Models, Retrieval Augmented Generation, Vector Database*



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR GAMBAR	xi
DAFTAR TABEL	xii
DAFTAR KODE	xiii
DAFTAR LAMPIRAN	xiv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	7
1.3 Batasan Masalah	7
1.4 Tujuan Penelitian	8
1.5 Manfaat Penelitian	8
1.6 Sistematika Penulisan	9
BAB 2 LANDASAN TEORI	10
2.1 Artificial Intelligence (AI)	10
2.2 Large Language Model	10
2.2.1 Vector Database	11
BAB 3 METODOLOGI PENELITIAN	17
3.1 Studi Literatur	17
3.2 Identifikasi Kebutuhan dan Permasalahan	17
3.2.1 Analisis Kebutuhan	17
3.3 Pemilihan Dataset	18
3.4 Perancangan Aplikasi	19
3.4.1 Rancangan RAG	19
3.4.2 Teknik populasi data Vector Database	24
3.5 Implementasi Teknik RAG Untuk LLM	33
3.6 Integrasi dengan Aplikasi	40
3.6.1 React JS untuk pengembangan frontend	40
3.6.2 Http Request	40
3.6.3 Flask sebagai Backend API	41
3.6.4 Desain mockup frontend	41
3.7 Pengujian dan Evaluasi	42
3.7.1 Metode Evaluasi	42
3.7.2 Penilaian	46
3.8 Dokumentasi (Penulisan Laporan)	47
3.9 Spesifikasi Sistem	47
3.9.1 Software	47
3.9.2 Hardware	48
BAB 4 HASIL DAN DISKUSI	49
4.1 Hasil Pengujian	49
4.1.1 Distance Metric	49

4.1.2	Chunk Retrieval Size	50
4.1.3	Embedding Model	51
BAB 5	SIMPULAN DAN SARAN	53
5.1	Simpulan	53
5.2	Saran	53
	DAFTAR PUSTAKA	55



UMMN

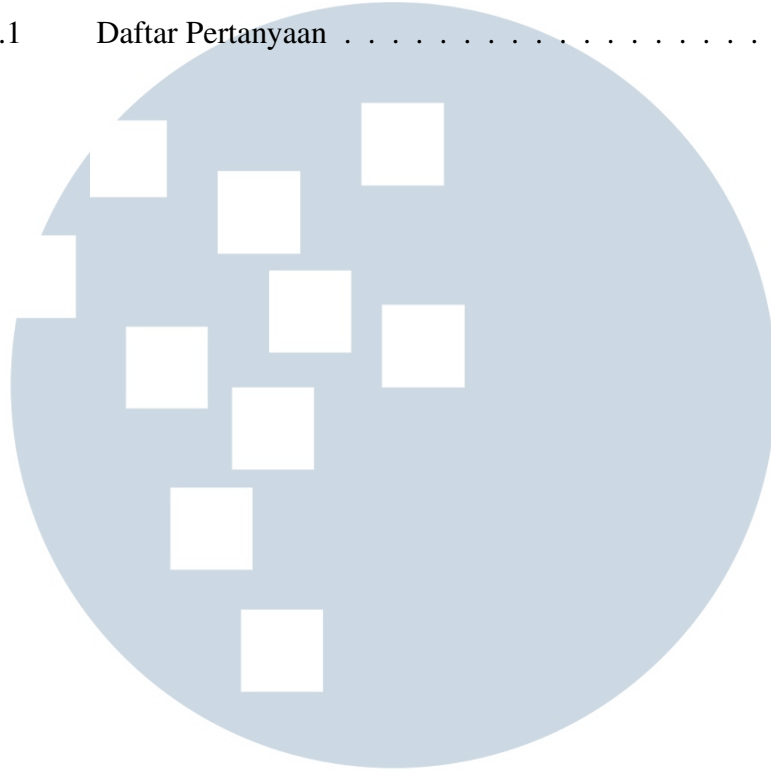
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

Gambar 2.1	Visualisasi data vector database	12
Gambar 2.2	Visualisasi query vector database	13
Gambar 2.3	Rangkaian proses implementasi Retrieval Augmented Generation	16
Gambar 3.1	End to end flowchart	20
Gambar 3.2	Flowchart RAG	21
Gambar 3.3	GPT Chat Completion Flowchart	22
Gambar 3.4	GPT Chat Completion Flowchart	23
Gambar 3.5	Teknik text splitting	24
Gambar 3.6	Source code proses read document	25
Gambar 3.7	Source code proses chunking document	25
Gambar 3.8	Inisiasi koneksi weaviate	26
Gambar 3.9	Inisiasi schema menggunakan distance metric manhattan	26
Gambar 3.10	Source code menyisipkan data kedalam vector store	27
Gambar 3.11	Document vector data	27
Gambar 3.12	Proses konversi teks menjadi vector data	28
Gambar 3.13	Proses indexing vector database HNSW graph	29
Gambar 3.14	Expectation Data Distribution	30
Gambar 3.15	Populate vector DB flowchart	31
Gambar 3.16	Read data from file flowchart	32
Gambar 3.17	Weaviate Role	33
Gambar 3.18	Weaviate Prompt	33
Gambar 3.19	Get Prompt Keywords	34
Gambar 3.20	Source code print vector information	34
Gambar 3.21	Prompt Keyword Generation	35
Gambar 3.22	Keyword vector representation	35
Gambar 3.23	Weaviate Semantic Search	35
Gambar 3.24	Hasil Semantic Search	35
Gambar 3.25	Greedy Search in Navigable Small World	37
Gambar 3.26	Hierarchical Navigable Small World Graph	38
Gambar 3.27	Chatbot Role Template	38
Gambar 3.28	Chatbot Prompt Template	39
Gambar 3.29	Source Code proses augmentasi dan generasi	39
Gambar 3.30	Source Code Pengiriman API menggunakan API OpenAI	39
Gambar 3.31	Hasil augmentasi dan generasi LLM	40
Gambar 3.32	Mockup Aplikasi	42
Gambar 3.33	Generate Question source code 1	43
Gambar 3.34	Generate Question source code 2	43
Gambar 3.35	Generate Question source code 3	43
Gambar 3.36	Source code proses generate actual output aplikasi RAG	46
Gambar 3.37	Source Code pengetesan G-Eval menggunakan DeepEval	47

DAFTAR TABEL

Tabel 3.1 Daftar Pertanyaan 44



UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR KODE

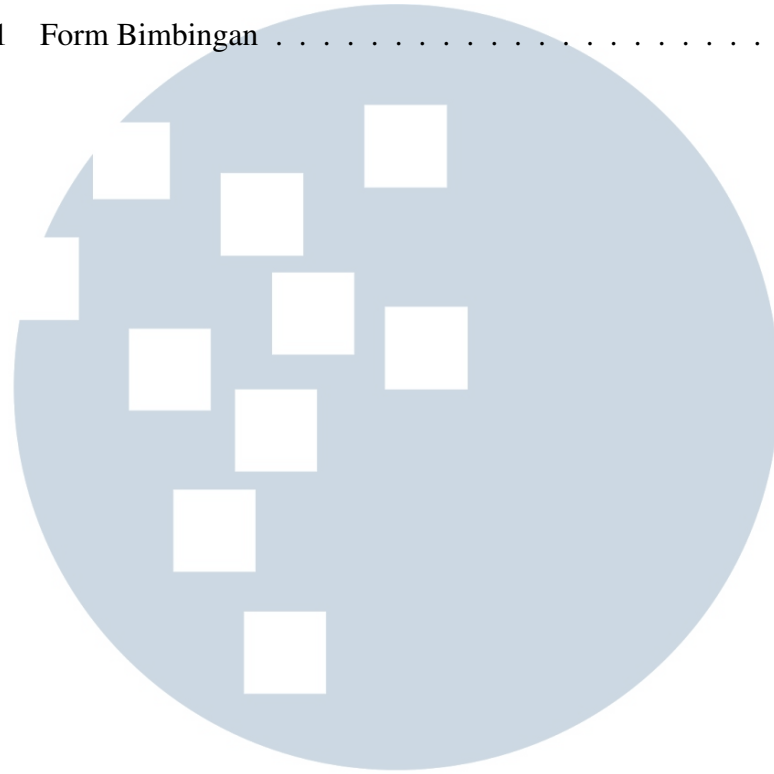


UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran 1 Form Bimbingan 57



UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA