

**Tesis**

Optimalisasi Arsitektur *Retrieval-Augmented Generation* melalui *Fine-Tuning Model Embedding* dan *Large Language Model* untuk Analisis  
*Nahwu-Sharaf* Kitab Kuning



# PENGESAHAN TUGAS AKHIR



KEMENTERIAN AGAMA  
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA  
FAKULTAS SAINS DAN TEKNOLOGI  
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

## PENGESAHAN TUGAS AKHIR

Nomor : B-1935/Un.02/DST/PP.00.9/08/2026

Tugas Akhir dengan judul : Optimalisasi Arsitektur Retrieval-Augmented Generation melalui Fine-Tuning Model Embedding dan Large Language Model untuk Analisis Nahwu-Sharaf Kitab Kuning

yang dipersiapkan dan disusun oleh:

Nama : SUBHAN GHUFRON, S.Kom  
Nomor Induk Mahasiswa : 24206051015  
Telah diujikan pada : Senin, 10 Agustus 2026  
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

### TIM UJIAN TUGAS AKHIR



Ketua Sidang

Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng.  
SIGNED

Valid ID: 6a86944493f1c



Penguji I

Dr. Agung Fatwanto, S.Si., M.Kom.  
SIGNED

Valid ID: 6a856831a4b9d



Penguji II

Muhammad Mustakim, S.T. M.T.  
SIGNED

Valid ID: 6a8687c70b2fa



Yogyakarta, 10 Agustus 2026  
UIN Sunan Kalijaga  
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.  
SIGNED

Valid ID: 6a8785c3e2e2b

## PERNYATAAN KEASLIAN

### PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Subhan Ghufron  
NIM : 24206051015  
Jenjang : Magister  
Program Studi : Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali pada bagian-bagian yang dirujuk sumbernya.

Yogyakarta, 10 Agustus 2026

Saya yang menyatakan,



**Subhan Ghufron**

NIM: 24206051015

STATE ISLAMIC UNIVERSITY  
**SUNAN KALIJAGA**  
YOGYAKARTA

## NOTA DINAS PEMBIMBING

### NOTA DINAS PEMBIMBING

Kepada Yth.,  
Dekan Fakultas Sains dan Teknologi  
UIN Sunan Kalijaga  
Yogyakarta

*Assalamu'alaikum wr. wb.*

Setelah melakukan bimbingan, arahan, dan koreksi terhadap penulisan tesis yang berjudul:  
*Optimalisasi Arsitektur Retrieval-Augmented Generation melalui Fine-Tuning Model Embedding dan Large Language Model untuk Analisis Nahwu-Sharaf Kitab Kuning*

Yang ditulis oleh:

|               |                  |
|---------------|------------------|
| Nama          | : Subhan Ghufron |
| NIM           | : 24206051015    |
| Jenjang       | : Magister       |
| Program Studi | : Informatika    |

Saya berpendapat bahwa tesis tersebut sudah dapat diajukan kepada Magister Informatika UIN Sunan Kalijaga untuk diujikan dalam rangka memperoleh gelar Magister Informatika.

*Wassalamu'alaikum wr. wb.*

Yogyakarta, 10 Agustus 2026

Pembimbing,



(Prof. Dr. Ir. Shofwatul 'Uyun, S.T.,  
M.Kom., IPM., ASEAN Eng.)

## ABSTRAK

Analisis gramatikal Arab (*Nahwu-Sharaf*) pada teks kitab kuning merupakan fondasi utama dalam tradisi pembelajaran pesantren di Indonesia. Proses ini masih bergantung pada analisis manual yang membutuhkan keahlian mendalam dan waktu yang lama. Penelitian ini bertujuan mengevaluasi efektivitas *fine-tuning* pada model *embedding* dan *Large Language Model* (LLM) dalam arsitektur *Retrieval-Augmented Generation* (RAG) untuk analisis *Nahwu-Sharaf* kitab kuning secara otomatis. Dua belas konfigurasi diuji secara komparatif, melibatkan dua model *embedding* (*BGE-M3* dan *E5-Large*) serta dua LLM (*Llama3-8B* dan *Mistral-7B*), masing-masing dalam kondisi murni dan setelah *fine-tuning* menggunakan teknik *QLoRA* pada *dataset NahwuAI*. Evaluasi dilakukan menggunakan metrik otomatis (*Recall@K*, *BLEU*, *ROUGE-L*, *F1*, *Cosine Similarity*). Hasil menunjukkan bahwa *BGE-M3 fine-tuned* unggul sebagai model *embedding* dengan *Recall@1* sebesar 0,8109 (naik 37,7%). *Mistral-7B fine-tuned* terpilih sebagai LLM terbaik karena kemampuan generalisasinya yang lebih kuat pada teks kitab kuning dibandingkan *Llama3-8B* yang cenderung *overfitting* terhadap data latih. Konfigurasi RAG + *Mistral fine-tuned* mencapai *F1* tertinggi (0,1176). RAG terbukti memberikan nilai tambah terbesar ketika dipasangkan pada model yang sudah kokoh melalui *fine-tuning*.

**Kata Kunci:** *Retrieval-Augmented Generation*, *Nahwu-Sharaf*, Kitab Kuning, *Fine-tuning*, Model *Embedding*, *Large Language Model*, *BGE-M3*, *E5-Large*, *Llama3-8B*, *Mistral-7B*

## ABSTRACT

*Arabic grammatical analysis (Nahwu-Sharaf) of classical Islamic texts (kitab kuning) is a cornerstone of the pesantren education tradition in Indonesia. This process still relies on manual analysis requiring deep expertise and extensive time. This study aims to evaluate the effectiveness of fine-tuning embedding models and Large Language Models (LLMs) within a Retrieval-Augmented Generation (RAG) architecture for automated Nahwu-Sharaf analysis of kitab kuning. Twelve configurations were tested comparatively, involving two embedding models (BGE-M3 and E5-Large) and two LLMs (Llama3-8B and Mistral-7B), each in baseline and fine-tuned conditions using QLoRA on the NahwuAI dataset. Evaluation was conducted using automatic metrics (Recall@K, BLEU, ROUGE-L, F1, Cosine Similarity) and human evaluation by three Arabic grammar experts. Results showed that fine-tuned BGE-M3 outperformed as the embedding model with a Recall@1 of 0.8109 (a 37.7% increase). Fine-tuned Mistral-7B was selected as the best LLM due to its stronger generalization capability on kitab kuning texts compared to Llama3-8B, which tended to overfit to training data. The RAG + fine-tuned Mistral configuration achieved the highest F1 (0.1176) and the highest expert score (3.80 out of 5). RAG was demonstrated to provide the greatest added value when paired with a model already strengthened through fine-tuning.*

**Keywords:** Retrieval-Augmented Generation, Nahwu-Sharaf, Kitab Kuning, Fine-tuning, Embedding Model, Large Language Model, BGE-M3, E5-Large, Llama3-8B, Mistral-7B

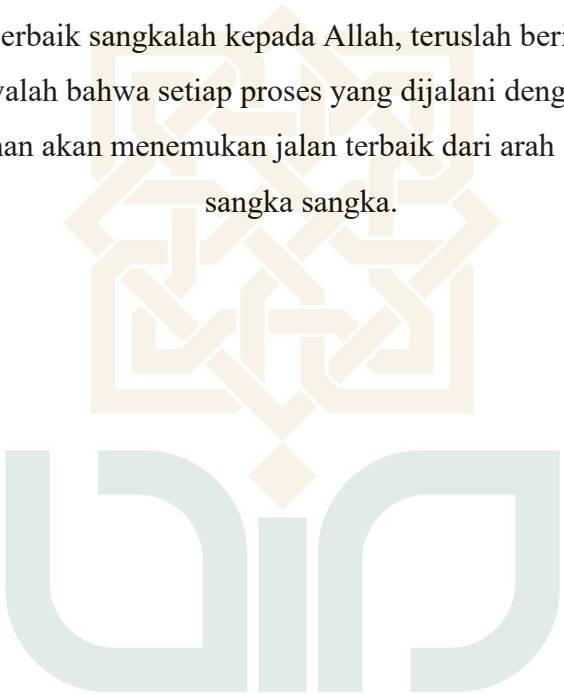
## MOTTO

أَنَا عِنْدَ ظَنِّ عَبْدِي بِي

*"Aku sesuai dengan prasangka hamba-Ku kepada-Ku."*

### **-Hadis Qudsi, Riwayat al-Bukhari dan Muslim-**

Maka, berbaik sangkalah kepada Allah, teruslah berikhtiar, dan percayalah bahwa setiap proses yang dijalani dengan penuh kesungguhan akan menemukan jalan terbaik dari arah yang tidak di sangka sangka.



STATE ISLAMIC UNIVERSITY  
SUNAN KALIJAGA  
YOGYAKARTA

## PERSEMBAHAN

Dengan penuh rasa syukur, tesis ini saya persembahkan kepada:

- Allah Swt., atas segala rahmat, pertolongan, dan kesempatan yang diberikan hingga karya ini dapat diselesaikan.
- Kedua orang tua dan keluarga tercinta, atas doa, kasih sayang, pengorbanan, dan dukungan yang tidak pernah berhenti mengiringi setiap langkah.
- Beasiswa PBSB (Program Beasiswa Santri Berprestasi)-LPDP Kementrian Agama RI, atas kesempatan dan kepercayaan yang diberikan untuk melanjutkan pendidikan. Saya yakin amanah ini dapat saya wujudkan melalui ilmu, karya, dan kebermanfaatan bagi masyarakat, pesantren, dan bangsa.
- Para guru, dosen, dan pembimbing, atas ilmu, bimbingan, nasihat, dan keteladanan yang menjadi bagian penting dalam perjalanan ini.
- Sahabat dan teman seperjuangan, atas doa, dukungan, kebersamaan, dan semangat yang menguatkan di setiap proses.
- Dan untuk diri sendiri, terima kasih telah bertahan, terus belajar, dan menyelesaikan setiap langkah dengan penuh kesungguhan.

Saya yakin karya sederhana ini menjadi wujud syukur dan dapat memberikan manfaat bagi sesama.

## KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah SWT atas rahmat dan karunia-Nya sehingga tesis dengan judul "Optimalisasi Arsitektur *Retrieval-Augmented Generation* melalui *Fine-Tuning Model Embedding* dan *Large Language Model* untuk Analisis *Nahwu-Sharaf Kitab Kuning*" dapat diselesaikan. Shalawat serta salam semoga tetap tercurah kepada Nabi Muhammad SAW, keluarga, sahabat, dan seluruh pengikutnya hingga akhir zaman.

Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister pada Program Studi Informatika, Program Magister, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga Yogyakarta. Penelitian ini membahas perbandingan model embedding dan LLM dalam arsitektur RAG untuk menganalisis Nahwu-Sharaf pada teks kitab kuning, dengan harapan dapat memberikan kontribusi bagi pengembangan teknologi NLP untuk bahasa Arab, khususnya dalam konteks pembelajaran di pesantren.

Penulis menyadari bahwa penyelesaian tesis ini tidak lepas dari bantuan, bimbingan, dan dukungan berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Prof. Noorhaidi, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor UIN Sunan Kalijaga Yogyakarta.
2. Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
3. Dr. Ir. Sumarsono, S.T., M.Kom., selaku Ketua Program Studi Magister Informatika UIN Sunan Kalijaga Yogyakarta.
4. Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng., selaku dosen pembimbing yang telah meluangkan waktu untuk

membimbing, memberi masukan, dan mengarahkan penulis sejak awal penyusunan proposal hingga tesis ini selesai.

5. Dr. Agung Fatwanto, S.Si., M.Kom dan Dr. Muhammad Mustakim, S.T. M.T., selaku dosen penguji yang telah memberikan koreksi dan masukan yang membuat tesis ini menjadi lebih baik.
6. Seluruh dosen Program Magister Informatika UIN Sunan Kalijaga Yogyakarta yang telah memberikan ilmu selama masa perkuliahan.
7. Program Beasiswa Santri Berprestasi (PBSB) - LPDP, yang telah membiayai seluruh proses studi penulis, sehingga penulis dapat menyelesaikan pendidikan magister ini tanpa kendala biaya.
8. Keluarga Besar Pondok Pesantren Al Falahiyyah yang telah memberikan dorongan dan motivasi untuk terus melanjutkan studi.
9. Bapak, Mamak dan seluruh keluarga penulis, yang selalu memberikan doa, dukungan, dan semangat selama proses studi ini berlangsung.
10. Semua pihak yang tidak dapat disebutkan satu per satu, yang telah membantu baik secara langsung maupun tidak langsung dalam penyelesaian tesis ini.

Penulis menyadari bahwa tesis ini masih memiliki banyak kekurangan, baik dari segi isi maupun penulisan. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan untuk perbaikan di masa mendatang. Semoga tesis ini dapat bermanfaat bagi pembaca,

khususnya bagi pengembangan penelitian di bidang NLP bahasa Arab dan pembelajaran kitab kuning.

Yogyakarta, 20 Agustus 2026

Penulis,

Subhan Ghufroon

NIM. 24206051015



## DAFTAR SINGKATAN

|        |                                                                                            |
|--------|--------------------------------------------------------------------------------------------|
| AI     | : <i>Artificial Intelligence</i>                                                           |
| BLEU   | : <i>Bilingual Evaluation Understudy</i>                                                   |
| BGE-M3 | : <i>BAAI General Embedding - Multi-Linguality, Multi-Functionality, Multi-Granularity</i> |
| GQA    | : <i>Grouped Query Attention</i>                                                           |
| GPU    | : <i>Graphics Processing Unit</i>                                                          |
| HNSW   | : <i>Hierarchical Navigable Small World</i>                                                |
| LLM    | : <i>Large Language Model</i>                                                              |
| LoRA   | : <i>Low-Rank Adaptation</i>                                                               |
| MAP    | : <i>Mean Average Precision</i>                                                            |
| MNRL   | : <i>Multiple Negatives Ranking Loss</i>                                                   |
| MRR    | : <i>Mean Reciprocal Rank</i>                                                              |
| MSA    | : <i>Modern Standard Arabic</i>                                                            |
| NDCG   | : <i>Normalized Discounted Cumulative Gain</i>                                             |
| NLP    | : <i>Natural Language Processing</i>                                                       |
| PEFT   | : <i>Parameter-Efficient Fine-tuning</i>                                                   |
| QLoRA  | : <i>Quantized Low-Rank Adaptation</i>                                                     |
| RAG    | : <i>Retrieval-Augmented Generation</i>                                                    |
| ROUGE  | : <i>Recall-Oriented Understudy for Gisting Evaluation</i>                                 |
| SWA    | : <i>Sliding Window Attention</i>                                                          |
| VRAM   | : <i>Video Random Access</i>                                                               |

## DAFTAR ISI

|                              |      |
|------------------------------|------|
| PENGESAHAN TUGAS AKHIR ..... | ii   |
| PERNYATAAN KEASLIAN .....    | iii  |
| NOTA DINAS PEMBIMBING .....  | iv   |
| ABSTRAK .....                | v    |
| <i>ABSTRACT</i> .....        | vi   |
| MOTTO .....                  | vii  |
| PERSEMBAHAN .....            | viii |
| KATA PENGANTAR.....          | ix   |
| DAFTAR SINGKATAN.....        | xii  |
| DAFTAR ISI .....             | xiii |
| DAFTAR TABEL .....           | xv   |
| DAFTAR GAMBAR.....           | xvi  |
| DAFTAR LAMPIRAN .....        | xvii |
| BAB I PENDAHULUAN .....      | 1    |
| A. Latar Belakang.....       | 1    |
| B. Rumusan Masalah .....     | 6    |
| C. Batasan Masalah .....     | 6    |
| D. Tujuan.....               | 8    |
| E. Manfaat.....              | 9    |
| BAB V PENUTUP .....          | 86   |

|                     |    |
|---------------------|----|
| A. Kesimpulan.....  | 86 |
| B. Saran .....      | 87 |
| DAFTAR PUSTAKA..... | 89 |



## DAFTAR TABEL

|                                                                                            |        |
|--------------------------------------------------------------------------------------------|--------|
| Tabel 2 1 Perbandingan sistematis antara penelitian ini dan studi-studi terkait. ....      | 21     |
| Tabel 3. 1 Spesifikasi Perangkat Keras Penelitian.....                                     | 34     |
| Tabel 3. 2 Daftar Perangkat Lunak Penelitian .....                                         | 35     |
| Tabel 3. 3 Skema penelitian.....                                                           | 37     |
| Tabel 3. 4 Ringkasan Model yang Digunakan.....                                             | 41     |
| Tabel 3. 5 Karakteristik Dataset NahwuAI.....                                              | 42     |
| Tabel 3. 6 Konfigurasi Hyperparameter Fine tuning Model Embedding<br>.....                 | 45     |
| Tabel 3. 7 Parameter Fine-tuning Model LLM.....                                            | 47     |
| Tabel 3. 8 Metrik Evaluasi Generation dan Implementasinya.....                             | 49     |
| <br>Tabel 4. 1 Konfigurasi Hyperparameter Fine-tuning Model Embedding<br>.....             | <br>52 |
| Tabel 4. 2 Perbandingan Keempat Konfigurasi Model Embedding ....                           | 55     |
| Tabel 4. 3 Konfigurasi Hyperparameter Fine-tuning QLoRA Mistral-<br>7B dan Llama3-8B ..... | 57     |
| Tabel 4. 4 Perbandingan Hasil Evaluasi Fine-Tuning Mistral-7B vs.<br>Llama3-8B .....       | 59     |
| Tabel 4. 5 Hasil Evaluasi Baseline B1 — Llama3-8B Murni (Kitab<br>Kuning) .....            | 61     |
| Tabel 4. 6 Hasil Evaluasi Baseline B2, Mistral-7B Murni (Kitab<br>Kuning) .....            | 62     |
| Tabel 4. 7 Hasil Evaluasi Baseline B3, Llama3-8B Fine-tuned (Kitab<br>Kuning) .....        | 62     |
| Tabel 4. 8 Hasil Evaluasi Baseline B4 — Mistral-7B Fine-tuned (Kitab<br>Kuning) .....      | 63     |

## DAFTAR GAMBAR

Gambar 3. 1 Alur Tahapan Penelitian . **Error! Bookmark not defined.**



## DAFTAR LAMPIRAN

Lampiran 1 Sample hasil dari RAG + Mistral Finetuned.....**Error!**

**Bookmark not defined.**

Lampiran 2 Sample hasil dari Baseline Llama 3 only.....**Error!**

**Bookmark not defined.**

Lampiran 3 Bukti Submit Jurnal Scopus.....**Error! Bookmark not defined.**



## BAB I

### PENDAHULUAN

#### A. Latar Belakang

Bahasa Arab merupakan salah satu bahasa yang memiliki kompleksitas gramatikal tinggi dan menjadi bahasa utama dalam tradisi keilmuan Islam. Di Indonesia, pembelajaran bahasa Arab terutama berlangsung di pesantren menggunakan kajian kitab kuning. Kitab kuning adalah literatur klasik berbahasa Arab yang memuat berbagai disiplin ilmu keislaman. Pemahaman gramatikal Arab yang meliputi *Nahwu* (sintaksis) dan *Sharaf* (morfologi) menjadi fondasi utama dalam memahami teks-teks klasik tersebut. Menurut (Aleyda, 2023) bahasa Arab memiliki karakteristik unik seperti sistem akar kata (*root system*), morfologi yang kaya, dan struktur sintaksis kompleks yang membedakannya dari bahasa-bahasa lain.

Analisis gramatikal Arab mencakup beberapa aspek penting, di antaranya adalah *i'rab* (perubahan akhir kata berdasarkan posisi dalam kalimat), *tarkib* (susunan kata dan struktur kalimat), serta kaidah-kaidah *Nahwu-Sharaf*. Penguasaan aspek-aspek ini membutuhkan keahlian mendalam dan waktu pembelajaran yang lama. Sebagaimana dicatat dalam (Al-sarayreh et al., n.d.) bahasa Arab menghadirkan tantangan unik bagi peneliti *Natural Language Processing* (NLP) karena struktur yang kompleks dan rumit mencakup variasi dialek, ambiguitas, dan struktur linguistik yang kaya.

Perkembangan teknologi kecerdasan buatan, khususnya *Large Language Models* (LLM), telah membuka peluang baru terutama dalam pemrosesan bahasa alami. LLM seperti *Llama 3* yang dikembangkan oleh Meta AI (Aaron Grattafiori, 2024) dan *Mistral 7B*

(Jiang et al., 2023) telah menunjukkan kemampuan yang luar biasa dalam berbagai tugas pemrosesan bahasa alami. Model-model ini memiliki kemampuan luar biasa dalam memahami konteks, menghasilkan teks yang koheren, dan melakukan *reasoning* yang kompleks.

Salah satu arsitektur yang menjanjikan untuk mengoptimalkan kemampuan LLM dalam domain spesifik adalah *Retrieval-Augmented Generation* (RAG). (Lewis et al., n.d.) memperkenalkan RAG sebagai pendekatan yang menggabungkan kekuatan *retrieval-based* dan *generation-based models* untuk meningkatkan kualitas generasi teks. Arsitektur RAG memungkinkan model untuk mengakses pengetahuan eksternal yang relevan sebelum menghasilkan respons, sehingga dapat meningkatkan akurasi dan mengurangi halusinasi pada domain yang memerlukan pengetahuan spesifik.

Penerapan RAG untuk bahasa Arab telah mulai dieksplorasi oleh beberapa peneliti. (El-Beltagy et al., 2024) melakukan studi komprehensif tentang implementasi dan evaluasi RAG untuk teks Arab. Penelitian tersebut mengeksplorasi berbagai model *semantic embedding* pada tahap *retrieval* dan beberapa LLM pada tahap *generation*. Hasil penelitian menunjukkan bahwa model *embedding* semantik yang ada dan LLM dapat digunakan secara efektif untuk membangun *pipeline* RAG bahasa Arab. Meskipun penelitian tersebut menunjukkan efektivitas *pipeline* RAG untuk teks Arab umum, penelitian tersebut belum secara khusus mengevaluasi teks kitab kuning yang memiliki karakteristik linguistik klasik, struktur sintaksis kompleks, dan terminologi keislaman yang khas. Keterbatasan ini membuka peluang bagi penelitian ini untuk mengisi celah tersebut dengan fokus pada

analisis *Nahwu-Sharaf* dalam teks-teks Arab klasik di lingkungan pesantren Indonesia

Dalam konteks *retrieval*, model *embedding* memainkan peran krusial. (Wang et al., 2024) mengembangkan *Multilingual E5*, model *embedding* teks *open-source* yang mendukung lebih dari 100 bahasa termasuk bahasa Arab. Model ini dilatih menggunakan *contrastive pre-training* pada 1 miliar pasangan teks *multilingual*. Sementara itu, (Chen et al., n.d.) memperkenalkan *M3-Embedding (BGE-M3)* yang unggul dalam *Multi-Linguality*, *Multi-Functionality*, dan *Multi-Granularity*. *BGE-M3* mampu melakukan *dense retrieval*, *multi-vector retrieval*, dan *sparse retrieval* secara bersamaan, serta dapat memproses *input* hingga 8.192 *token*.

Meskipun telah ada kemajuan dalam penerapan AI untuk pemrosesan bahasa Arab, masih terdapat beberapa kesenjangan penelitian yang signifikan. Pertama, kitab kuning belum sepenuhnya dianalisis secara otomatis menggunakan teknologi modern. Kedua, analisis manual gramatikal Arab membutuhkan keahlian mendalam dan waktu yang lama, sehingga membatasi aksesibilitas pembelajaran. Ketiga, Kajian yang secara khusus mengevaluasi performa LLM dalam analisis *i'rab* dan *tarkib* pada teks kitab kuning masih sangat terbatas, terutama yang mengintegrasikan mekanisme *retrieval* berbasis *semantic embedding*. Keempat, belum ada penelitian yang secara khusus membandingkan pengaruh *fine-tuning* pada model *embedding* dan LLM untuk analisis gramatikal Arab dalam arsitektur RAG.

Fine-tuning penuh pada model berukuran miliaran parameter menuntut kapasitas VRAM yang sulit dipenuhi oleh perangkat keras tunggal. Untuk mengatasi kendala tersebut, (Dettmers et al., 2023)

memperkenalkan Quantized Low-Rank Adaptation (QLoRA), yaitu metode fine-tuning yang membekukan bobot model pre-trained dalam representasi terkuantisasi 4-bit NormalFloat (NF4) dan menyisipkan adapter Low-Rank Adaptation (LoRA) yang dapat dilatih pada setiap layer Transformer. Efisiensi memori diperkuat oleh Double Quantization dan Paged Optimizers, sehingga fine-tuning model berukuran 65 miliar parameter dapat dilakukan pada satu GPU 48 GB tanpa menurunkan kualitas dibandingkan fine-tuning presisi penuh 16-bit. Pendekatan ini relevan untuk fine-tuning model embedding dan LLM pada domain Nahwu-Sharaf dengan sumber daya komputasi terbatas.

Ketersediaan *dataset NahwuAI* yang dapat diakses melalui *Hugging Face* (robinsyihab/*NahwuAi-dataset*) membuka peluang untuk melakukan *fine-tuning* model secara spesifik untuk domain analisis gramatikal Arab. *Dataset* ini berisi pasangan pertanyaan-jawaban tentang kaidah *Nahwu-Sharaf* yang dapat digunakan untuk melatih model agar lebih memahami konteks gramatikal Arab. Dengan demikian, penelitian ini sangat urgen untuk dilakukan guna mengembangkan sistem *chatbot* berbasis RAG yang dapat menganalisis gramatikal Arab *Nahwu-Sharaf* dengan akurat, serta memberikan penjelasan dalam bahasa Indonesia yang dapat dipahami oleh santri dan pelajar.

Analisis Nahwu-Sharaf termasuk tugas yang memerlukan pengetahuan khusus. Setiap kalimat Arab yang dianalisis membutuhkan rujukan pada kaidah tertentu, misalnya aturan i'rab atau syarat muftada-khabar. Kaidah-kaidah tersebut tersebar di banyak kitab Nahwu klasik dan tidak selalu terwakili di data pelatihan LLM umum. Llama 3 dan

Mistral dilatih pada teks web berskala besar yang didominasi bahasa Inggris dan Arab modern, sedangkan analisis Nahwu-Sharaf membutuhkan pengetahuan Arab klasik yang lebih spesifik. Oleh karena itu, LLM yang hanya mengandalkan pengetahuan internalnya rentan menghasilkan analisis yang keliru atau mengarang jawaban.

Karakteristik kitab kuning memperbesar tantangan tersebut. Teks kitab kuning ditulis tanpa harakat, sehingga satu bentuk tulisan bisa punya beberapa kemungkinan bacaan dengan makna gramatikal yang berbeda. Kosakata klasik yang dipakai juga jarang muncul di teks Arab modern. Kondisi tersebut membuat LLM tanpa dukungan retrieval semakin mudah menghasilkan halusinasi, yaitu jawaban yang terdengar meyakinkan tetapi salah secara kaidah. Retrieval-Augmented Generation (RAG) menjadi pendekatan yang tepat untuk mengatasi masalah ini. (Lewis et al.,) menunjukkan bahwa penggabungan komponen retrieval dan generation dapat menekan halusinasi karena LLM diberi konteks pengetahuan yang relevan sebelum menghasilkan jawaban. (El-Beltagy et al., 2024) memperkuat temuan tersebut dengan membuktikan bahwa pipeline RAG efektif untuk memproses teks Arab menggunakan LLM *open-source*. Berdasarkan pertimbangan tersebut, arsitektur RAG dipilih sebagai pendekatan yang paling sesuai untuk analisis Nahwu-Sharaf pada teks kitab kuning.

Berdasarkan latar belakang tersebut, penelitian ini akan melakukan Optimalisasi Arsitektur *Retrieval-Augmented Generation* melalui *Fine-Tuning Model Embedding* dan *Large Language Model* untuk Analisis Nahwu-Sharaf Kitab Kuning. Penelitian ini diharapkan dapat menemukan kombinasi model terbaik yang dapat menghasilkan analisis *i'rab* dan *tarkib* yang akurat, serta memberikan kontribusi pada

pengembangan sistem AI untuk pembelajaran gramatikal Arab di Indonesia

## B. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan diatas, maka rumusan masalah dalam penelitian ini adalah: " Bagaimana optimalisasi performa model *embedding* dan *Large Language Model* yang telah di-*fine-tuning* dalam sistem *Retrieval-Augmented Generation* untuk analisis *Nahwu-Sharaf* pada teks kitab kuning? "

1. Bagaimana perbandingan performa model *embedding E5-Large* dan *BGE-M3*, baik sebelum maupun sesudah *fine-tuning*, dalam menghasilkan *retrieval* yang relevan untuk *query* analisis gramatikal Arab?
2. Bagaimana perbandingan performa model *Llama 3* dan *Mistral*, baik sebelum maupun sesudah *fine-tuning*, dalam menghasilkan analisis *i'rab* dan *tarkib* yang akurat pada teks kitab kuning?
3. Bagaimana perbandingan performa sistem RAG dengan pendekatan *LLM-only* tanpa mekanisme *retrieval* dalam analisis *Nahwu-Sharaf* pada teks kitab kuning?

## C. Batasan Masalah

Untuk menjaga fokus dan ruang lingkup penelitian agar tetap terarah, maka batasan penelitian ini ditetapkan sebagai berikut:

Berikut batasan penelitian yang lebih tepat, disusun berdasarkan apa yang dibatasi dan konsekuensinya terhadap generalisasi hasil, bukan sekadar mendata apa yang dilakukan:

1. Penelitian ini dibatasi pada analisis gramatikal tingkat kalimat (*i'rab* dan *tarkib*), sehingga aspek *balaghah*, tafsir makna, maupun analisis wacana lintas-kalimat tidak dicakup.
2. Teks uji dibatasi pada kitab kuning berbahasa Arab klasik (*fusha turats*) dengan struktur matan-syarah. Dengan demikian, hasil penelitian tidak dapat digeneralisasi untuk bahasa Arab modern standar (MSA), dialek *amiyah*, maupun teks Arab kontemporer di luar tradisi pesantren.
3. Perbandingan model dibatasi pada dua model *embedding* (*multilingual-E5-large*, *bge-m3*) dan dua LLM (*Llama-3-8B*, *Mistral-7B-v0.3*) yang dipilih berdasarkan pertimbangan kesetaraan ukuran parameter dan ketersediaan lisensi sumber terbuka. Model *embedding* dan LLM lain di luar kedua pasangan ini termasuk model komersial seperti GPT atau Gemini berada di luar cakupan perbandingan.
4. Penyempurnaan model dibatasi pada pendekatan parameter-efficient fine-tuning dengan QLoRA (kuantisasi 4-bit NF4 + adapter LoRA), sehingga temuan penelitian ini tidak mencerminkan potensi hasil dari full fine-tuning maupun pelatihan model dari tahap pretraining..
5. *Dataset fine-tuning* dibatasi pada *NahwuAI Dataset*. Cakupan kaidah *Nahwu-Sharaf* dalam penelitian ini karena itu terbatas pada variasi kaidah yang termuat dalam *dataset* tersebut, dan tidak mewakili seluruh mazhab maupun aliran tata bahasa Arab klasik yang ada.

6. Keluaran sistem dibatasi pada penjelasan gramatikal berbahasa Indonesia, sehingga kesesuaian sistem untuk pengguna dengan latar bahasa lain tidak diuji.
7. Penelitian ini tidak mencakup pengujian pada kitab kuning di luar sampel data uji yang digunakan, sehingga performa sistem pada judul kitab, penulis, atau mazhab fikih-*Nahwu* lain di luar sampel belum tentu setara.

#### D. Tujuan

Secara umum tujuan dari penelitian ini adalah mengevaluasi efektivitas *fine-tuning* pada model *embedding* dan *Large Language Model* dalam sistem *Retrieval-Augmented Generation* (RAG) untuk analisis *Nahwu-Sharaf* pada teks kitab kuning..

Adapun tujuan khusus dari penelitian ini adalah sebagai berikut:

1. Mengoptimalkan performa model *embedding E5-Large* dan *BGE-M3*, baik sebelum maupun sesudah *fine-tuning*, dalam menghasilkan *retrieval* yang relevan untuk analisis *Nahwu-Sharaf*.
2. Mengoptimalkan performa model *Llama 3* dan *Mistral*, baik sebelum maupun sesudah *fine-tuning*, dalam menghasilkan analisis *i'rab* dan *tarkib* pada teks kitab kuning..
3. Mengevaluasi perbandingan performa sistem berbasis RAG dan pendekatan LLM- *only* dalam analisis *Nahwu-Sharaf* pada teks kitab kuning.

4. Melakukan eksperimen komparatif metode *embedding* terbaik dan model *generation* untuk menganalisis *tarkib* bahasa Arab dalam kitab kuning.
5. Mengevaluasi seluruh metode menggunakan metrik otomatis (*Recall@K*, *Precision@K*, *MRR*, *BLEU*, *F1*, *Cosine Similarity*) dan evaluasi manusia (*human evaluation*).

#### E. Manfaat

Studi ini diharapkan memberikan manfaat baik secara teoritis maupun praktis sebagai berikut:

1. Manfaat teoritis
  - a. Kontribusi ilmiah di bidang Pemrosesan Bahasa Alami (NLP), khususnya dalam pemrosesan bahasa Arab menggunakan arsitektur RAG.
  - b. Menyediakan studi komparatif yang komprehensif mengenai pengaruh *fine-tuning* terhadap model *embedding* dan LLM untuk pengetahuan domain-spesifik, khususnya tata bahasa Arab.
  - c. Memperkaya literatur mengenai penerapan AI untuk analisis linguistik klasik, dengan fokus pada tradisi ilmiah *Nahwu-Sharaf* dalam teks-teks Arab klasik.
2. Manfaat praktis
  - a. Bagi santri : Menyediakan alat bantu untuk memahami analisis tata bahasa Arab dalam “Kitab Kuning” dengan penjelasan yang mudah dipahami dalam bahasa Indonesia.

- b. Bagi guru dan ustadz: Mendukung proses pembelajaran *Nahwu-Sharaf* melalui sistem yang secara otomatis dan konsisten memberikan penjelasan mengenai *I'rab* dan *Tarkib*.
- c. Bagi pengembang AI: Penyediaan tolok ukur dan praktik terbaik untuk pengembangan sistem NLP (Pemrosesan Bahasa Alami) untuk bahasa Arab, khususnya dalam konteks analisis tata bahasa.
- d. Bagi pesantren dan lembaga pendidikan Islam: Mendukung upaya digitalisasi dan modernisasi pembelajaran “Kitab Kuning”, tanpa kehilangan esensi tradisi keilmuan klasik.

## BAB V

### PENUTUP

#### A. Kesimpulan

Penelitian ini menjawab pertanyaan yang menjadi acuan sejak awal, yaitu sistem berbasis *Retrieval-Augmented Generation* seperti apa yang paling baik menganalisis *Nahwu-Sharaf* kitab kuning ke dalam bahasa Indonesia. Dari dua belas konfigurasi yang diuji, satu konfigurasi keluar sebagai pemenang yang konsisten di seluruh jalur evaluasi, yaitu RAG + *Mistral fine-tuned* yang memadukan *BGE-M3 fine-tuned* sebagai pencari dokumen dengan *Mistral-7B fine-tuned* sebagai penjawab.

Tiga pertanyaan penelitian terjawab dengan angka yang saling searah. Untuk sisi pencarian dokumen, *BGE-M3* yang dilatih ulang dengan *dataset NahwuAI* menaikkan *Recall@1* dari 0,5889 ke 0,8109 (naik 37,7 persen) dan *MRR* dari 0,6859 ke 0,8640 (naik 25,96 persen). *BGE-M3 fine-tuned* unggul atas *E5-Large* di semua level K, sehingga dipakai untuk seluruh percobaan Fase 2. Untuk sisi *generator*, *Mistral-7B fine-tuned* mengalahkan *Llama3-8B fine-tuned* baik tanpa RAG maupun dengan RAG. F1 *Mistral fine-tuned* tanpa RAG mencapai 0,1021, sedangkan *Llama3 fine-tuned* tanpa RAG hanya 0,0944. Pasangan RAG-nya, yaitu RAG + *Mistral fine-tuned* mencapai 0,1176 sedangkan RAG + *Llama3 fine-tuned* 0,1041.

Hasil paling menarik untuk dicatat datang dari benturan antara hasil pada data latihan dan hasil pada kitab kuning nyata. Pada *validation set NahwuAI*, *Llama3 fine-tuned* terlihat lebih kuat hampir di semua metrik. Begitu model bertemu teks kitab kuning yang gaya

bahasanya berbeda dari data latihan, posisi berbalik tajam. *Cosine Llama3* turun 28,7 persen, *ROUGE-L* turun 68,3 persen, dan *BLEU-4* nyaris habis di 98,9 persen. Kemampuan *Mistral* hampir tidak berubah sama sekali, *Cosine* bahkan naik tipis 0,7 persen. Pola ini punya pesan metodologis yang jelas. Nilai bagus di data latihan tidak boleh dijadikan satu-satunya rujukan untuk memilih model yang akan dipakai di lapangan. Pelatihan ulang dan RAG juga bukan dua jalur yang bisa saling menggantikan, tapi jalur yang saling melengkapi, dan paling produktif ketika pasangannya cocok.

Hasil dari metrik evaluasi tersebut belum sepenuhnya baik dan sesuai dengan harapan. Angka-angka tersebut masih terbilang sangat rendah untuk analisis *nahwu shorof*. Penggunaan RAG dalam kasus ini juga belum bisa memberikan efek yang maksimal. Justru *finetuning* yang memberikan efek lebih besar dari pada RAG. Maka dari itu masih diperlukan pengembangan dan riset lanjutan agar keakuratannya menjadi lebih baik.

## B. Saran

Bagi peneliti lanjutan di area *Nahwu-Sharaf*, pelajaran paling jelas dari penelitian ini ada di sisi data. *Dataset NahwuAI* sudah cukup baik untuk pelatihan berbasis instruksi, tapi belum cukup untuk membentuk kemampuan menalar kaidah secara mendalam. Penelitian berikutnya dapat memperkaya data latihan dengan contoh-contoh yang memuat penjelasan lebih lengkap untuk tiap kaidah, bukan hanya pasangan pertanyaan-jawaban singkat. Sumber bacaan yang dipakai pencari dokumen juga akan lebih berdampak kalau ditambah teks kitab kuning klasik secara langsung, sehingga contoh yang ditarik dekat dengan gaya

bahasa teks yang sedang dianalisis. Perlu dilakukan eksperimen terhadap model LLM lain seperti Qwen, Komodo, AraBert dan sebagainya untuk menemukan model yang paling tepat.



## DAFTAR PUSTAKA

- Abdelazim, H. (2023). Semantic Embeddings for Arabic Retrieval Augmented Generation ( ARAG ). *International Journal of Advanced Computer Science and Applications*, 14(11).
- Al-ghamdi, S., & Al-khalifa, H. (2024). Fine-Tuning BERT-Based Pre-Trained Models for Arabic Dependency Parsing. *Applied Sciences*, 14(17). <https://doi.org/10.3390/app14177654>
- Al-sarayreh, S., Mohamed, A., & Shaalan, K. (n.d.). *Challenges and Solutions for Arabic Natural Language Processing in Social*. Springer Nature Singapore. <https://doi.org/10.1007/978-981-99-3416-4>
- Alan, M. F., Karagoz, P., & Toroslu, I. H. (2024). MufassirQAS: A Qur'anic Question Answering System Based on RAG and Vector Database. *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*.
- Aleyda, M. A. br G. (2023). Uniknya Bahasa Arab. *Ekhsis: Jurnal Ekonomi, Syariah Dan Studi Islam*, 1(2).
- Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools* (pp. 9–15). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.wanlp-1.1>
- Bhatia, G., Nagoudi, E. M. B., El Mekki, A., Alwajih, F., & Abdul-Mageed, M. (2025). Swan and ArabicMTEB: Dialect-Aware, Arabic-Centric, Cross-Lingual, and Cross-Cultural Embedding Models and Benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

- <https://doi.org/10.18653/v1/2025.findings-naacl.263>
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2318–2335). Association for Computational Linguistics.  
<https://doi.org/10.18653/v1/2024.findings-acl.992>
- Darwish, K., Habash, N., Abbas, M., Al-khalifa, H., Al-natsheh, H. T., El-beltagy, S. R., Bouamor, H., Bouzoubaa, K., Cavalli-sforza, V., & El-hajj, W. (2021). A Panoramic Survey of Natural Language Processing in the Arab World. *Communications of the ACM*, 64(4), 72–81. <https://doi.org/10.1145/3448247>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized Language Models. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.48550/arXiv.2305.14314>
- El-Beltagy, S. R., Abdelali, A., & Mubarak, H. (2024). Exploring Retrieval Augmented Generation in Arabic. *Procedia Computer Science*, 244, 296–307. <https://doi.org/10.1016/j.procs.2024.10.203>
- El-sheikh, A., Elmogtaba, A., Darwish, K., Elmallah, M., Elneima, A., & Sawaf, H. (n.d.). *Creating Arabic LLM Prompts at Scale*. 3–8.
- Fauzi, A., Hamid, A., & Nurdin, N. (2025). Development of STEM-Literacy-Based Me-Book for Arabic Language Learning in Pesantren. *International Journal of Arabic Language Teaching*.
- Foundation, A., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen,

- Z., Mohammed, O., Samta, A., Shen, Z., Liu, Z., Vassilieva, N., Hestness, J., Hock, A., Feldman, A., Lee, J., Jackson, A., Xuguang, H., Nakov, P., ... Xing, E. (2023). Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *Arxiv*.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., & Sun, J. (2024). Retrieval-Augmented Generation for Large Language Models : A Survey. *Arxiv*, 1–21.
- Grattafiori, A., & al., et. (2024). The Llama 3 Herd of Models. *ArXiv:2407.21783*. <https://doi.org/10.48550/arXiv.2407.21783>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=nZeVKeeFYf9>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. *ArXiv Preprint*, *abs/2310.06825*. <https://doi.org/10.48550/arXiv.2310.06825>
- Koubaa, A., Nacar, O., & Sibae, S. (2025a). Enhanced Arabic Retrieval Augmented Generation Using Nested Embedding Models. In *Smart Systems and Emerging Technologies* (pp. 120–131). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-81521-4\\_10](https://doi.org/10.1007/978-3-031-81521-4_10)
- Koubaa, A., Nacar, O., & Sibae, S. (2025b). *Enhanced Arabic*

- Retrieval Augmented Generation Using Nested Embedding Models BT - International Conference on Smart Systems and Emerging Technologies* (A. Koubaa, A. Ben Mnaouer, W. Boulila, & S. Raghay (eds.); pp. 120–131). Springer Nature Switzerland.
- Kwon, S. J., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction. In *Proceedings of the 6th Workshop on Arabic NLP (ArabicNLP 2023)*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.arabiclelp-1.9>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*. <https://doi.org/10.5555/3495724.3496517>
- Mahmoud, Z., Li, C., Zappatore, M., Solyman, A., Alfatemi, A., Ibrahim, A. O., & Abdelmaboud, A. (2023). *Semi-supervised learning and bidirectional decoding for effective grammar correction in low-resource scenarios*. 1–25. <https://doi.org/10.7717/peerj-cs.1639>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. <https://doi.org/10.3115/1073083.1073135>
- Rini, R., Mustofa, A., & Kurniawan, R. (2022). Transformation of Arabic Learning from Classical Model to Digital Model. *Ijaz Arabi*

*Journal of Arabic Learning*, 5(3).

<https://doi.org/10.18860/ijazarabi.v5i3.17409>

Sopian, A., Abdurahman, M., Ali, M., Tantowi, Y. A., Aeni, A. N., & Maulani, H. (2025). Arabic Language Learning in a Multicultural Context at Pesantren. *Jurnal Pendidikan Islam*, 11(1).  
<https://doi.org/10.15575/jpi.v11i1.44104>

Space, O. I. (2024). *Building an Arabic RAG Pipeline: Integration of GATE-AraBERT, ARA-Reranker, and GPT-4 Mini*. Hugging Face / GitHub. <https://huggingface.co/Omartificial-Intelligence-Space>

Team, F., Abbas, U., Ahmad, M. S., Alam, F., & al., et. (2025). Fanar: An Arabic-Centric Multimodal Generative AI Platform. *ArXiv:2501.13944*. <https://doi.org/10.48550/arXiv.2501.13944>

Vaswani, A. (2017). *Attention Is All You Need*. *Nips*.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *ArXiv:2402.05672*. <https://doi.org/10.48550/arXiv.2402.05672>

Zbib, R. (2025). Aftina: A RAG-Enhanced System for Islamic Fatwa Generation. In *Proceedings of ArabicNLP*.

STATE ISLAMIC UNIVERSITY  
 SUNAN KALIJAGA  
 YOGYAKARTA