Received: April 2026	Accepted: September 2026	Published: Oktober 2026

Pengembangan Sistem Tanya Jawab untuk Pencarian Hadis Tematik Menggunakan Model Llama-3 dengan Pendekatan *Retrieval-Augmented Generation*

Zannuba[1*], Hendri Ahmadian [2], Nurrizqa[3]

[1*,2,3] *Teknologi Informasi, Sains dan Teknologi, Universitas Islam Negeri Ar-Raniry Banda Aceh*

[*]*Email: zannubanuba@gmail.com*

Abstract

Thematic hadith retrieval remains challenging because hadith collections contain numerous narrations distributed across various books and topics, while users often do not know the exact wording or source of the hadith they seek. This study aims to develop a thematic hadith retrieval system based on Retrieval-Augmented Generation (RAG) using Llama 3.1-8B-Instruct, adapted through fine-tuning with Low-Rank Adaptation (LoRA). The system utilizes 33.738 hadiths from six major collections, BAAI/bge-m3 to generate semantic representations, and FAISS to retrieve the three most relevant hadiths. The retrieved hadiths are used as context for the model to generate brief explanations. The system presents the Arabic text, Indonesian translation, explanation, grade, and reference. Evaluation on 50 queries achieved a Hit Rate@3 of 1.0000 and an NDCG@3 of 0.9435. Answer quality evaluation using RAGAS showed that the fine-tuned + RAG model achieved a Faithfulness score of 0.8523 and an Answer Relevancy score of 0.5439, which were higher than those of the base model + RAG, with scores of 0.8285 and 0.5394, respectively. These results indicate that the system consistently retrieves relevant hadiths, while LoRA fine-tuning improves the quality of answers generated based on the retrieved context, although relevance filtering and control over model elaboration still require further improvement.

Keywords: Thematic hadith retrieval, large language model, llama 3.1, retrieval-augmented generation, low-rank adaptation

Abstrak

Pencarian hadis berdasarkan tema masih menghadapi kendala karena koleksi hadis mencakup banyak riwayat dalam berbagai kitab dan topik, sedangkan pengguna sering tidak mengetahui redaksi maupun sumber hadis yang dicari. Penelitian ini bertujuan mengembangkan sistem pencarian hadis tematik berbasis *Retrieval-Augmented Generation* (RAG) menggunakan Llama 3.1-8B-Instruct yang diadaptasi melalui *fine-tuning* dengan teknik *Low-Rank Adaptation* (LoRA). Sistem memanfaatkan 33.738 hadis dari enam kitab utama, BAAI/bge-m3 untuk membentuk representasi semantik, serta FAISS untuk memperoleh tiga hadis paling relevan. Hadis hasil *retrieval* digunakan sebagai konteks bagi model untuk menghasilkan penjelasan singkat. Sistem menyajikan teks Arab, terjemahan, penjelasan, derajat, dan referensi. Evaluasi terhadap 50 kueri menghasilkan Hit Rate@3 sebesar 1.0000 dan NDCG@3 sebesar 0.9435. Evaluasi kualitas jawaban menggunakan RAGAS menunjukkan bahwa model *fine-tuned* + RAG memperoleh *Faithfulness* sebesar 0.8523 dan *Answer Relevancy* sebesar 0.5439, lebih tinggi dibandingkan *base model* + RAG yang memperoleh 0.8285 dan 0.5394. Hasil tersebut menunjukkan bahwa sistem mampu menemukan hadis relevan secara konsisten, sedangkan *fine-tuning* LoRA memberikan peningkatan pada kualitas jawaban berdasarkan konteks hasil *retrieval*, meskipun penyaringan relevansi dan pembatasan elaborasi model masih perlu disempurnakan.

Kata kunci: *Pencarian hadis tematik, large language model, llama 3.1, retrieval-augmented generation, low-rank adaptation*

1. Pendahuluan

Hadis merupakan sumber hukum Islam kedua setelah Al-Qur'an yang memiliki peran penting sebagai landasan dalam pembentukan akidah, ibadah, akhlak, serta berbagai aspek kehidupan umat Islam [1]. Koleksi hadis dalam berbagai kitab utama, seperti *Sahih al-Bukhari, Sahih Muslim, Sunan Abu Dawud, Sunan an-Nasa'i, Jami' at-Tirmidhi*, dan *Sunan Ibn Majah*, mencakup puluhan ribu riwayat yang disusun berdasarkan beragam bab dan tema. Dalam proses penelusuran, pengguna dapat mengajukan kebutuhan informasi berdasarkan suatu tema atau permasalahan, seperti sedekah, kejujuran, dan adab bertetangga, tanpa mengetahui letak kitab maupun redaksi asli riwayat. Besarnya jumlah hadis menjadikan proses pencarian hadis yang relevan terhadap suatu permasalahan tidak selalu mudah dilakukan secara manual. Oleh karena itu, ketersediaan informasi hadis yang akurat, mudah diakses, dan dapat dipertanggungjawabkan menjadi semakin penting untuk mendukung pemahaman dan praktik keagamaan masyarakat [2].

Perkembangan kecerdasan buatan, khususnya *Large Language Model* (LLM), telah mendorong pengembangan sistem tanya jawab berbasis bahasa alami yang mampu memahami pertanyaan pengguna dan menghasilkan respons secara kontekstual [3]. Kemampuan tersebut membuka peluang pemanfaatan LLM dalam pengembangan sistem pencarian hadis. Namun, LLM masih rentan menghasilkan informasi yang tampak benar dan meyakinkan, tetapi tidak didukung oleh sumber yang valid. Fenomena ini dikenal sebagai *hallucination* dan menjadi salah satu tantangan utama dalam pemanfaatan LLM pada sistem tanya jawab [4]. Pada domain hadis, permasalahan tersebut berpotensi menyebabkan kesalahan redaksi, atribusi kitab yang tidak tepat, maupun penyajian hadis yang tidak memiliki dasar otoritatif. Kondisi ini menunjukkan perlunya pendekatan yang mampu menghasilkan jawaban kontekstual sekaligus mempertahankan keterkaitannya dengan sumber hadis yang dapat diverifikasi.

Salah satu pendekatan yang dapat digunakan untuk mengatasi keterbatasan tersebut adalah *Retrieval-Augmented Generation* (RAG). Pendekatan ini menggabungkan mekanisme *retrieval* untuk memperoleh dokumen relevan dari basis pengetahuan dengan kemampuan generatif LLM dalam menyusun jawaban berdasarkan informasi yang berhasil diambil [5]. Dengan menjadikan dokumen hasil *retrieval* sebagai konteks, RAG dapat meningkatkan akurasi dan kredibilitas jawaban serta mengurangi kemungkinan model menghasilkan informasi yang tidak didukung oleh sumber. Selain itu, penggunaan dokumen eksternal juga membuat dasar informasi dalam jawaban lebih mudah ditelusuri dan diverifikasi oleh pengguna [6].

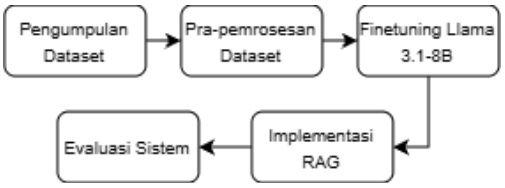
RAG telah diterapkan pada berbagai sistem tanya jawab keagamaan dengan cakupan yang beragam. Alan dkk. [7] mengembangkan MufassirQAS dengan mengintegrasikan RAG berbasis FAISS dan ChatGPT-3.5 Turbo. Ahadi dkk. [8] membangun sistem tanya jawab fikih mazhab Syafi'i menggunakan Llama 3.2:3B dengan pendekatan RAG dan memperoleh nilai F1 BERTScore sebesar 0.8552. Pada domain hadis, Herwanza dkk. [9] dan Syah dkk. [10] mengembangkan sistem berbasis LangChain Retriever dan GPT-4-1106-preview dengan hasil evaluasi yang baik, tetapi masih bergantung pada model bahasa tertutup. Penelitian lain menunjukkan bahwa model generatif tanpa adaptasi domain dapat menghasilkan interpretasi yang lebih luas dibandingkan model ekstraktif yang telah disesuaikan dengan data keagamaan [11]. Khalila dkk. [12] mengevaluasi 13 model bahasa terbuka yang dipadukan dengan RAG pada domain studi Al-Qur'an dan menemukan bahwa kemampuan model memengaruhi relevansi serta kualitas jawaban. Penerapan RAG pada studi Al-Qur'an dan sistem tanya jawab Islam berbahasa Persia juga menunjukkan bahwa kualitas jawaban dipengaruhi oleh kemampuan model, proses pengambilan konteks, dan karakteristik bahasa yang digunakan [13].

Berdasarkan kajian tersebut, penerapan RAG pada sistem pencarian hadis masih menyisakan ruang untuk pengembangan lebih lanjut. Penelitian pada domain hadis masih banyak memanfaatkan model bahasa tertutup atau proprietari, seperti GPT-4-1106-preview, sehingga fleksibilitas dalam penyesuaian model menjadi terbatas [9], [10]. Sementara itu, pada penelitian yang menggunakan model berbobot terbuka (*open-weight model*) maupun pendekatan berbasis LLM pada domain keislaman [12], [13], integrasi model bahasa berbobot terbuka dengan pendekatan RAG untuk mendukung pencarian hadis tematik berbahasa Indonesia masih belum menjadi fokus utama dalam penelitian-penelitian yang dikaji.

Penelitian ini mengembangkan sistem pencarian hadis tematik berbasis RAG menggunakan model Llama 3.1-8B-Instruct yang diadaptasi melalui *fine-tuning* dengan teknik *Low-Rank Adaptation* (LoRA). Sistem memanfaatkan mekanisme *retrieval* untuk memperoleh hadis yang relevan terhadap pertanyaan pengguna. Selanjutnya, model menghasilkan penjelasan singkat berdasarkan isi setiap hadis yang ditemukan. Hasil pencarian disajikan dalam bentuk teks Arab, terjemahan, penjelasan singkat, derajat, dan referensi hadis sehingga sumber informasi dapat diperiksa kembali oleh pengguna. Melalui integrasi model hasil *fine-tuning* dan RAG, sistem diharapkan dapat mendukung pencarian hadis tematik yang lebih relevan, kontekstual, dan transparan.

2. Metode Penelitian

Penelitian ini menggunakan metode eksperimen untuk mengembangkan sistem pencarian hadis tematik berbasis *Retrieval-Augmented Generation* (RAG) dengan memanfaatkan model Llama 3.1-8B yang diadaptasi melalui *fine-tuning* dengan teknik *Low-Rank Adaptation* (LoRA). Untuk mendukung proses *retrieval*, setiap dokumen hadis direpresentasikan dalam bentuk *embedding* menggunakan model BAAI/bge-m3 dan disusun ke dalam indeks vektor menggunakan FAISS. Tahapan penelitian meliputi pra-pemrosesan dataset, *fine-tuning* model, pembangunan basis pengetahuan, implementasi mekanisme RAG, serta evaluasi kinerja sistem. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1. Pengumpulan Data

Dataset penelitian ini menggunakan repositori meeAtif/hadith_datasets yang tersedia secara publik pada Hugging Face Datasets. Dataset tersebut dipilih karena mencakup 33.738 hadis dari enam kitab utama, sebagaimana ditunjukkan pada Tabel 1. Setiap entri dataset memuat informasi berupa teks hadis dalam bahasa Arab dan bahasa Inggris, judul bab, derajat hadis (Grade), nama kitab (Book), serta referensi hadis (Reference). Atribut Reference menyediakan tautan ke Sunnah.com untuk penelusuran sumber hadis, sedangkan atribut lainnya digunakan dalam pengolahan data untuk sistem pencarian hadis tematik.

Tabel 1. Distribusi Dataset per Kitab Hadis

No.	Nama Kitab	Jumlah Entri
1	Sahih al-Bukhari	7.563
2	Sahih Muslim	7.190
3	Sunan Abu Dawud	5.274
4	Sunan an-Nasa'i	5.758
5	Jami' at-Tirmidhi	3.956
6	Sunan Ibn Majah	3.997
	Total	33.738

2.2. Pra-pemrosesan Data

2.2.1 Translasi Dataset

Teks hadis berbahasa Inggris diterjemahkan ke dalam bahasa Indonesia menggunakan pustaka *deep-translator* yang dijalankan pada Google Colab agar sistem dapat mendukung pertanyaan pengguna berbahasa Indonesia. Setelah proses translasi, dilakukan pemeriksaan secara terbatas terhadap beberapa sampel hasil terjemahan untuk

memastikan kelengkapan dan keterbacaan hasil terjemahan. Namun, pemeriksaan kesesuaian makna secara menyeluruh terhadap seluruh hasil translasi dan validasi oleh ahli bahasa atau ahli hadis belum dilakukan.

2.2.2 Data Cleaning

Data hasil translasi dibersihkan melalui normalisasi *whitespace* dengan menghapus karakter *newline* dan spasi berlebih pada kolom teks Arab dan terjemahan menggunakan ekspresi reguler.

2.2.3 Penyusunan Dataset Instruction-Tuning

Dari total 33.738 hadis yang tersedia, penyusunan dataset *instruction-tuning* diawali dengan pengambilan sampel sebanyak 14–15 hadis dari masing-masing 320 bab, sehingga diperoleh 4.666 entri hadis. Selanjutnya, model Llama 3.1-8B-Instruct digunakan untuk menghasilkan penjelasan singkat sebagai respons. Sebelum dilakukan augmentasi, data dibagi dengan rasio 90:10 menjadi data training dan validation menggunakan metode stratified split berdasarkan judul bab. Pembagian dilakukan sebelum augmentasi agar variasi pertanyaan yang berasal dari hadis yang sama tidak tersebar ke kedua subset, sehingga risiko data *leakage* dapat diminimalkan. Hasil pembagian menghasilkan 4.199 hadis untuk training dan 467 hadis untuk validation. Setiap hadis kemudian diaugmentasi menjadi enam variasi pertanyaan, sehingga diperoleh 25.194 pasangan *instruction-response* untuk training dan 2.802 pasangan untuk validation, dengan total 27.996 pasangan data.

2.3. Fine-Tuning Model Llama 3.1-8B

Pada penelitian ini, *fine-tuning* dilakukan untuk mengadaptasi model pra-latih terhadap karakteristik domain hadis dan pola respons yang digunakan dalam sistem. Proses ini tidak ditujukan untuk menjadikan parameter model sebagai sumber utama pengetahuan faktual, karena informasi hadis tetap diperoleh melalui mekanisme *retrieval* pada RAG. Model dasar yang digunakan adalah Llama-3.1-8B-Instruct (melalui framework Unsloth) [14], yang di *fine-tune* menggunakan teknik LoRA (*Low-Rank Adaptation*) pada presisi 4-bit (quantized, fp16). LoRA merupakan salah satu teknik Parameter-Efficient *Fine-Tuning* (PEFT) yang membekukan parameter model pra-latih dan menambahkan matriks berdimensi rendah pada lapisan-lapisan tertentu sebagai parameter yang dilatih [15]. Pendekatan ini menghasilkan jumlah parameter yang diperbarui jauh lebih sedikit dibandingkan *fine-tuning* penuh, tanpa penurunan kualitas hasil yang signifikan. Adapter LoRA diterapkan pada seluruh modul proyeksi atensi dan feed-forward. Rincian konfigurasi arsitektur LoRA disajikan pada Tabel 2.

Tabel 2. Konfigurasi model

Parameter	Nilai
Model dasar	*unsloth/Llama-3.1-8B-Instruct*
Total parameter model	8.072.204.288
Kuantisasi	4-bit
Max Sequence Length	2.048 token
LoRA *rank* (r)	16
LoRA Alpha	16
Target modul	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Trainable parameters	41.943.040 (0.52% trained)

Dari Tabel 2 dapat dilihat bahwa konfigurasi LoRA menggunakan rank r=16 dengan alpha=16 pada tujuh modul target arsitektur *Transformer*, sehingga hanya 0.52% dari total parameter model dasar yang dilatih ulang selama proses adaptasi. Strategi ini secara signifikan menekan kebutuhan memori dan komputasi dibandingkan full *fine-tuning*, sehingga *fine-tuning* model 8B tetap dapat dilakukan pada lingkungan Google Colab dengan sumber daya terbatas. Selain konfigurasi arsitektur LoRA, proses pelatihan itu sendiri juga memerlukan pengaturan parameter tersendiri seperti yang dirangkum pada Tabel 3.

Tabel 3. Konfigurasi *Fine-tuning*

Parameter	Nilai
Jumlah Sampel *Fine-Tuning*	25.194
Epochs	1
Gradient Accumulation Steps	4
Total step	3.150
Learning rate	1×10^{-4}

Optimizer	AdamW 8-bit
Warmup Ratio	0.03
Weight Decay	0.05

Berdasarkan Tabel 3, *fine-tuning* dilakukan pada 25.194 sampel selama 1 *epoch* dengan 3.150 *training steps*, *gradient accumulation* sebesar 4, *learning rate* 1×10^{-4}, dan optimizer AdamW 8-bit. *Warmup ratio* sebesar 0.03 dan *weight decay* sebesar 0.05 digunakan untuk mendukung kestabilan proses pelatihan.

2.4. Implementasi *Retrieval-Augmented Generation* (RAG)

Untuk menghasilkan jawaban yang benar-benar berlandaskan hadis otentik dan menekan risiko halusinasi, penelitian ini mengintegrasikan pendekatan *Retrieval-Augmented Generation* (RAG) ke dalam model hasil *fine-tuning*. Sistem terlebih dahulu mengambil hadis yang relevan dari basis pengetahuan sebelum menyusun jawaban, sehingga tidak hanya mengandalkan pengetahuan yang tersimpan pada parameter model. Arsitektur sistem RAG pada penelitian ini disajikan pada Gambar 2.

Gambar 2. Alur Pipeline RAG

Sistem RAG pada penelitian ini terdiri atas dua tahap utama: tahap pengindeksan (dilakukan sekali di awal) dan tahap kueri (dijalankan setiap kali pengguna mengajukan pertanyaan). Pada tahap pengindeksan, seluruh korpus hadis yang berjumlah 33.738 hadis diubah menjadi representasi vektor menggunakan model *embedding* BAAI/bge-m3 (*multilingual dense embedding*), kemudian disimpan ke dalam indeks FAISS untuk mendukung pencarian kemiripan (*similarity search*) yang efisien.

Pada tahap kueri, pertanyaan pengguna terlebih dahulu di-encode dengan model *embedding* yang sama, lalu dicari tiga dokumen (*top-k*, k=3) paling relevan dari indeks FAISS. Kandidat hasil *retrieval* kemudian disaring melalui tahap filter relevansi berbasis LLM sebelum diteruskan ke tahap generasi. Hadis yang lolos filter diproses oleh model Llama-3.1-8B hasil *fine-tuning* untuk menghasilkan jawaban akhir dalam format terstruktur (Hadis Arab, Terjemahan, Penjelasan, Derajat, Referensi). Khusus untuk teks Arab, konten diambil langsung dari dokumen sumber (bukan hasil generasi model), karena Llama-3.1-8B ditemukan cenderung merusak karakter Arab bila dijadikan bagian dari keluaran generatif.

2.5. Evaluasi Sistem RAG

Evaluasi kinerja sistem dilakukan menggunakan dua kelompok metrik, yaitu kualitas *retrieval* dan kualitas jawaban.

2.5.1 Evaluasi Retrieval

Evaluasi *retrieval* bertujuan mengukur seberapa baik sistem menemukan hadis yang relevan dari indeks FAISS. Dua metrik digunakan adalah *Hit Rate@k* dan *Normalized Discounted Cumulative Gain* (NDCG@k) [16], [17].

a. Hit Rate@k

Hit Rate@k mengukur proporsi kueri yang berhasil mendapatkan minimal satu hadis relevan di antara hasil *retrieval*.

$$\text{Hit Rate@k} = \frac{1}{N}\sum_{i=1}^{N} \text{hit}_i\ @k$$

N menyatakan jumlah total kueri evaluasi, sedangkan hit pada kueri ke-i bernsilai 1 apabila terdapat minimal satu hadis relevan di antara hasil *retrieval* kueri tersebut, dan 0 apabila tidak ada. Nilai *Hit Rate@k* berada pada rentang 0 hingga 1, di mana semakin tinggi nilainya maka semakin konsisten sistem

menyertakan hadis relevan dalam tiga hasil *retrieval*.

b. NDCG@k

NDCG@k digunakan untuk mengukur kualitas urutan hasil *retrieval* dengan mempertimbangkan posisi hadis yang relevan dalam *k* hasil teratas. DCG@k merepresentasikan kualitas urutan hasil *retrieval* yang diperoleh sistem, sedangkan IDCG@k merupakan nilai DCG pada urutan ideal, yaitu ketika hadis-hadis relevan ditempatkan pada posisi teratas. NDCG@k dihitung dengan membandingkan DCG@k dengan IDCG@k sebagai berikut:

$$NDCG@K = \frac{DCG@k}{\text{IDCG@k}}$$

Nilai NDCG@k berada pada rentang 0 hingga 1. Nilai yang semakin mendekati 1 menunjukkan bahwa hadis relevan semakin banyak ditempatkan pada posisi atas dalam hasil retrieval.

2.5.2 Evaluasi Generation

Evaluasi generation dilakukan menggunakan framework RAGAS untuk menilai kualitas jawaban yang dihasilkan oleh sistem berdasarkan konteks hadis yang diperoleh dari proses *retrieval*. Penelitian ini menggunakan dua metrik, yaitu *Faithfulness* dan *Answer Relevancy* [18].

a. *Faithfulness*

Faithfulness mengukur sejauh mana klaim dalam jawaban yang dihasilkan model dapat ditelusuri kembali ke konteks yang di-*retrieve*, sebagai indikator deteksi halusinasi.

$$\text{Faithfulness} = \frac{\text{Jumlah klaim yang didukung oleh konteks}}{\text{Total klaim yang dihasilkan}}$$

Jawaban yang dihasilkan model terlebih dahulu dipecah menjadi klaim-klaim atomik oleh LLM judge, kemudian setiap klaim diperiksa apakah dapat disimpulkan dari konteks yang tersedia. Nilai *Faithfulness* berada pada rentang 0 hingga 1, di mana semakin tinggi nilainya maka semakin sedikit klaim dalam jawaban yang tidak berdasar pada hadis yang di-retrieve.

b. Answer Relevancy

Answer Relevancy mengukur relevansi semantik antara jawaban yang dihasilkan dan pertanyaan asli pengguna.

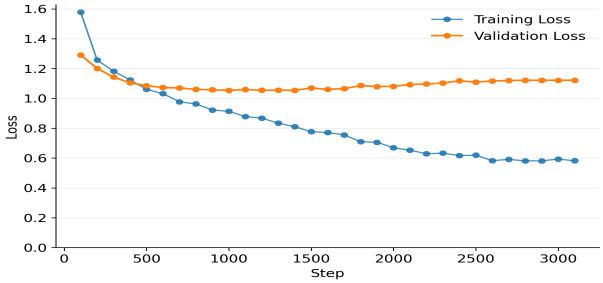
$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N} \text{sim}(q,q_i)$$

q menyatakan pertanyaan asli pengguna, sedangkan q dengan indeks i adalah salah satu dari N pertanyaan hasil rekonstruksi LLM berdasarkan jawaban yang dihasilkan sistem. Fungsi sim menghitung cosine similarity antara *embedding* keduanya. Nilai *Answer Relevancy* berada pada rentang 0 hingga 1, di mana semakin tinggi nilainya maka semakin fokus jawaban model terhadap apa yang sebenarnya ditanyakan.

3. Hasil dan Pembahasan

3.1. Hasil Training

Proses *fine-tuning* model dilakukan menggunakan metode *supervised fine-tuning* dengan adapter LoRA berdasarkan konfigurasi yang telah dijelaskan pada Tabel 2 dan Tabel 3. Model dilatih menggunakan 25.194 pasangan *instruction-response* sebagai data *training* dan 2.802 pasangan sebagai data *validation*. Selama proses pelatihan, nilai *training loss* dan *validation loss* dipantau untuk mengamati kemampuan model dalam mempelajari pola data sekaligus mengevaluasi kemampuan generalisasinya terhadap data yang tidak digunakan dalam proses pembaruan parameter. Proses pelatihan berlangsung selama satu *epoch* dengan total 3.150 *step*. Perkembangan nilai *training loss* dan *validation loss* ditunjukkan pada Gambar 3.



Gambar 3. Grafik *Training Loss* dan *Validation Loss*

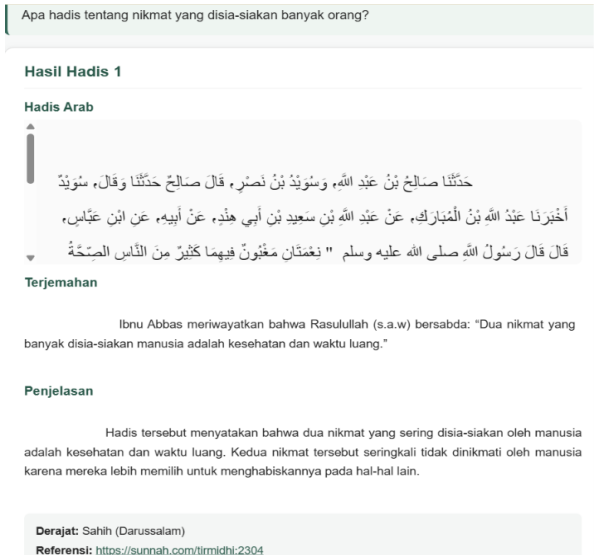
Gambar 3 menunjukkan perkembangan *training loss* dan *validation loss* selama proses *fine-tuning* yang dilakukan selama satu epoch dengan total 3.150 *step*. Pada tahap awal pelatihan, *training loss* dan *validation loss* sama-sama mengalami penurunan. *Training loss* menurun dari 1.5785 pada *step* 100 menjadi 0.9136 pada *step* 1000, sedangkan *validation loss* menurun dari 1.2910 menjadi 1.0539 pada rentang yang sama. Nilai *validation loss* terendah diperoleh pada *step* 1000, yaitu sebesar 1.0539. Setelah titik tersebut, *validation loss* relatif stabil hingga sekitar *step* 1400, kemudian secara umum menunjukkan kecenderungan meningkat hingga mencapai 1.1215 pada *step* 3100. Sebaliknya, *training loss* terus menurun hingga 0.5817 pada *step* 3100. Perbedaan kecenderungan antara kedua nilai *loss* tersebut menunjukkan bahwa setelah sekitar *step* 1400, model semakin menyesuaikan diri terhadap data pelatihan, sementara kinerja pada data validasi tidak lagi mengalami peningkatan.

Pelatihan dibatasi menjadi satu epoch karena model yang digunakan, yaitu Llama-3.1-8B-Instruct, merupakan model pralatih (*pre-trained model*) yang diadaptasi menggunakan LoRA. Dengan demikian, proses *fine-tuning* tidak bertujuan melatih model dari awal, melainkan menyesuaikan model terhadap karakteristik domain hadis dan format jawaban yang digunakan dalam penelitian. Selain itu, hasil pada Gambar 3 menunjukkan bahwa *validation loss* telah mencapai nilai minimum dalam satu epoch dan cenderung meningkat pada tahap berikutnya. Oleh karena itu, penambahan jumlah epoch berpotensi memperbesar *overfitting* tanpa memberikan perbaikan pada performa validasi. Selama pelatihan, checkpoint dengan nilai *validation loss* terendah dipilih sebagai model terbaik dan digunakan sebagai model akhir.

3.2. Hasil Implementasi RAG

Implementasi RAG pada penelitian ini digunakan untuk menghubungkan proses pencarian hadis dengan tahap pembentukan jawaban oleh model. Ketika pengguna mengajukan kueri, sistem mengambil tiga hadis yang paling relevan (k=3) dari basis data beserta informasi peringkat, jarak kemiripan, referensi sumber, tingkat kesahihan, teks Arab, dan terjemahan hadis. Hasil *retrieval* ini kemudian digunakan sebagai konteks pada tahap *generation*.

Berdasarkan konteks hasil *retrieval*, sistem menghasilkan jawaban yang ditampilkan kepada pengguna berupa teks Arab hadis beserta terjemahan, penjelasan singkat mengenai kandungan hadis, derajat kesahihan, dan referensi sumber, sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Contoh Hasil *Generation* Akhir Sistem

3.3. Hasil Evaluasi

Evaluasi sistem dilakukan terhadap 50 kueri pengujian yang disusun secara manual dengan mempertimbangkan keberagaman topik, mencakup berbagai domain pembahasan hadis seperti akidah, ibadah, akhlak, muamalah, thaharah, keluarga, larangan & dosa besar, serta lingkungan. Setiap kueri dilengkapi dengan data referensi berupa beberapa hadis relevan yang dijadikan acuan penilaian kinerja sistem. Evaluasi mencakup dua aspek utama, yaitu evaluasi *retrieval* dan evaluasi *generation*, yang masing-masing dijelaskan pada sub-bagian berikut.

3.3.1 Hasil Evaluasi Retrieval

Evaluasi kinerja sistem *retrieval* dilakukan menggunakan dua metrik utama, yaitu *Hit Rate@3* dan *Normalized Discounted Cumulative Gain (NDCG@3)*. Hasil evaluasi secara keseluruhan ditunjukkan pada Tabel 4.

Tabel 4. Hasil Evaluasi *Retrieval*

No.	Metrik	Skor rata-rata
1	Hit Rate@3	1.0000
2	NDCG@3	0.9435

Nilai Hit Rate@3 sebesar 1.0000 menunjukkan bahwa sistem berhasil menemukan setidaknya satu hadis yang relevan dalam tiga hasil teratas untuk seluruh 50 kueri pengujian. Metrik ini tidak mensyaratkan seluruh hasil relevan berada dalam top-3, melainkan cukup minimal satu dokumen relevan yang ditemukan.

Nilai NDCG@3 sebesar 0.9435 mencerminkan kualitas peringkat dokumen relevan secara keseluruhan dalam tiga posisi teratas. Nilai ini tergolong sangat tinggi, namun belum mencapai nilai sempurna. Berdasarkan analisis per kueri, penurunan ini disebabkan oleh beberapa kueri yang memiliki lebih dari satu hadis relevan dalam data referensi, sehingga tidak seluruh hadis relevan tersebut berhasil terjangkau dalam tiga hasil teratas. Sebagai contoh, kueri mengenai keutamaan sahur dan keutamaan zakat fitrah masing-masing hanya berhasil menemukan satu dari beberapa hadis relevan yang ada dalam data referensi, sehingga menghasilkan nilai NDCG@3 sebesar 0.47 pada kedua kueri tersebut.

Evaluasi *retrieval* tidak dibandingkan antara model dasar dan model hasil *fine-tuning* karena kedua konfigurasi menggunakan mekanisme *retrieval* yang identik, yaitu model *embedding* BAAI/bge-m3, indeks FAISS, dan nilai *top-k* sebesar 3. Proses *fine-tuning* LoRA hanya diterapkan pada model Llama-3.1-8B-Instruct pada tahap *generation*, sehingga tidak memengaruhi proses pengambilan maupun pemeringkatan hadis. Oleh karena itu, perbandingan kedua model difokuskan pada evaluasi *generation*.

3.3.2 Hasil Evaluasi Generation

Evaluasi dilakukan menggunakan framework RAGAS untuk mengukur dua metrik yaitu *Faithfulness* dan *Answer Relevancy*. Sebagai judge LLM digunakan GPT-5.4-mini melalui OpenAI API, terpisah dari model Llama yang di *fine-tune* sebagai model utama sistem.

Tabel 5. Hasil Evaluasi Fine-tuned Llama-3.1-8B+RAG

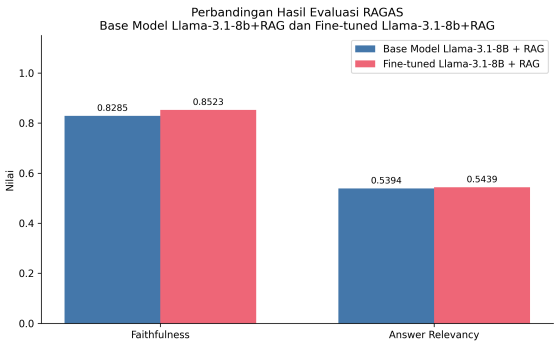
No.	Metrik	Skor rata-rata
1	Faithfulness	0.8523
2	Answer Relevancy	0.5439

Tabel 5 menampilkan hasil evaluasi sistem tanya jawab menggunakan model Llama-3.1-8B-Instruct hasil *fine-tuning* yang diintegrasikan dengan pendekatan RAG. Nilai *Faithfulness* sebesar 0.8523 menunjukkan bahwa jawaban yang dihasilkan memiliki tingkat keterdukungan yang tinggi terhadap konteks hadis hasil *retrieval*, sehingga sebagian besar informasi dalam jawaban tetap berlandaskan pada konteks yang diberikan. Sementara itu, nilai *Answer Relevancy* sebesar 0.5439 menunjukkan bahwa kesesuaian jawaban terhadap pertanyaan pengguna masih perlu ditingkatkan. Sebagai pembanding, evaluasi juga dilakukan pada sistem pencarian hadis tematik berbasis RAG yang menggunakan model dasar Llama-3.1-8B-Instruct tanpa *fine-tuning*. Perbandingan ini dilakukan untuk mengetahui pengaruh proses *fine-tuning* LoRA terhadap kualitas jawaban berdasarkan nilai *Faithfulness* dan *Answer Relevancy*

Tabel 6. Hasil Evaluasi Base Model Llama-3.1-8B+RAG

No.	Metrik	Skor rata-rata
1	Faithfulness	0.8285
2	Answer Relevancy	0.5394

Berdasarkan Tabel 6, hasil evaluasi *base model* Llama-3.1-8B-Instruct yang dikombinasikan dengan RAG memperoleh nilai *Faithfulness* sebesar 0.8285 dan *Answer Relevancy* sebesar 0.5394. Hasil evaluasi ini digunakan sebagai *baseline* untuk membandingkan kinerja model sebelum dan sesudah proses *fine-tuning*.



Gambar 5. Perbandingan Hasil Evaluasi

Berdasarkan hasil perbandingan, model hasil fine-tuning dengan RAG menunjukkan peningkatan pada kedua metrik dibandingkan base model dengan RAG. Nilai Faithfulness meningkat dari 0.8285 menjadi 0.8523, dengan peningkatan sebesar 0.0238, sedangkan Answer Relevancy meningkat dari 0.5394 menjadi 0.5439, dengan peningkatan sebesar 0.0045. Peningkatan tersebut menunjukkan bahwa proses fine-tuning LoRA membantu model menyesuaikan pola pembentukan jawaban berdasarkan konteks hadis yang diberikan oleh sistem RAG.

Karena informasi faktual tetap diperoleh melalui mekanisme retrieval, kontribusi *fine-tuning* pada penelitian ini lebih diarahkan pada adaptasi pola respons daripada penyimpanan pengetahuan hadis di dalam parameter model. Selain membandingkan kinerja model sebelum dan sesudah *fine-tuning*, dilakukan *error analysis* untuk mengidentifikasi jenis kesalahan yang masih muncul pada hasil evaluasi sistem. Analisis ini mencakup kesalahan pada tahap *retrieval* dan *generation*, sebagaimana dirangkum pada Tabel 7.

Tabel 7. *Error Analysis*

Jenis Kesalahan	Analisis
Elaborasi di luar konteks	Model menambahkan penjelasan yang tidak dinyatakan secara eksplisit dalam konteks terjemahan hadis sehingga menurunkan *Faithfulness*.
Ketidaksesuaian detail kueri	Hadis masih satu tema, tetapi perumpamaannya berbeda. Contoh: pada kueri "satu tubuh", satu hadis memakai perumpamaan "bangunan yang saling menopang" (HR. Tirmidzi No. 1928).
Keterbatasan cakupan *top-k*	Tidak semua hadis relevan masuk tiga hasil teratas. Contoh: kueri keutamaan sahur dan zakat fitrah masing-masing hanya memperoleh satu hadis relevan (NDCG@3 0.47).

4. Kesimpulan

Penelitian ini berhasil mengembangkan sistem pencarian hadis tematik berbasis Retrieval-Augmented Generation (RAG) menggunakan model Llama 3.1-8B-Instruct yang diadaptasi melalui fine-tuning dengan teknik Low-Rank Adaptation (LoRA). Sistem mampu menemukan hadis yang relevan terhadap pertanyaan pengguna serta menyajikannya dalam bentuk teks Arab, terjemahan, penjelasan singkat, derajat, dan referensi hadis. Berdasarkan pengujian terhadap 50 kueri, kinerja retrieval memperoleh nilai Hit Rate@3 sebesar 1.0000 dan NDCG@3 sebesar 0.9435. Hasil tersebut menunjukkan bahwa sistem secara konsisten menemukan setidaknya satu hadis relevan pada tiga hasil teratas dan secara umum mampu menempatkan hadis-hadis relevan pada urutan yang tinggi, meskipun pada beberapa kueri belum seluruh hadis relevan tercakup dalam tiga hasil teratas.

Evaluasi kualitas jawaban menggunakan RAGAS menunjukkan bahwa model hasil *fine-tuning* + RAG memperoleh nilai Faithfulness sebesar 0.8523 dan Answer Relevancy sebesar 0.5439. Sebagai pembanding, base model + RAG memperoleh nilai Faithfulness sebesar 0.8285 dan Answer Relevancy sebesar 0.5394. Dengan demikian, proses *fine-tuning* meningkatkan *Faithfulness* sebesar 0.0238 dan *Answer Relevancy* sebesar 0.0045. Hasil tersebut menunjukkan bahwa *fine-tuning* LoRA memberikan peningkatan terhadap kualitas jawaban, terutama dalam membantu model menyesuaikan pola pembentukan respons berdasarkan konteks hadis yang diperoleh

melalui proses *retrieval*. Meskipun demikian, hasil *error analysis* menunjukkan masih terdapat beberapa kesalahan berupa elaborasi di luar konteks, ketidaksesuaian hadis terhadap detail kueri, serta keterbatasan cakupan top-k. Dengan demikian, integrasi *fine-tuning* LoRA dan RAG mampu mendukung pencarian hadis tematik secara relevan, kontekstual, dan transparan, namun masih memerlukan penyempurnaan pada proses penyaringan hadis dan pembentukan jawaban.

5. Saran

Penelitian selanjutnya dapat menerapkan mekanisme reranking atau filter relevansi yang lebih ketat, memperkuat batasan pada *prompt* generasi, memperluas jumlah dan variasi kueri pengujian, serta melibatkan ahli bahasa atau ahli hadis dalam validasi hasil translasi dan kualitas hasil jawaban sistem.

6. Daftar Pustaka

[1] M. Asgari-Bidhendi, M. A. Ghaseminia, A. Shahbazi, S. A. Hossayni, N. Torabian, and B. Minaei-Bidgoli, "Rezwan: Leveraging Large Language Models for Comprehensive Hadith Text Processing: A 1.2M Corpus Development," 2025, [Online]. Available: http://arxiv.org/abs/2510.03781

[2] S. Susanti, I. Najiyah, Y. Ramdhani, A. Herliana, M. K. Muckti, and F. R. Oktaviani, "Searching Sahih Hadiths Based on Queries using Neural Models and FastText," *J. Appl. Data Sci.*, vol. 6, no. 1, pp. 272–285, 2025, doi: 10.47738/jads.v6i1.467.

[3] M. Yue, "A Survey of Large Language Model Agents for Question Answering," 2025, [Online]. Available: http://arxiv.org/abs/2503.19213

[4] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, 2023, doi: 10.1145/3571730.

[5] K. Martineau, A. I. Explainable, and A. I. Generative, "What is retrieval-augmented generation?," *IBM Res. Blog*, vol. 22, 2023.

[6] T. T. Procko and O. Ochoa, "Graph Retrieval-Augmented Generation for Large Language Models: A Survey," *Proc. - 2024 Conf. AI, Sci. Eng. Technol. AIxSET 2024*, pp. 166–169, 2024, doi: 10.1109/AIxSET62544.2024.00030.

[7] A. Y. Alan, E. Karaarslan, and O. Aydin, "Improving LLM Reliability with RAG in Religious Question-Answering: MufassirQAS," vol. 9, no. 3, pp. 544–559, 2025.

[8] R. Ahadi, N. Safaat Harahap, M. Fikry, and F. Kurnia, "Retrieval-Augmented Generation in a Web-Based Question Answering System for Fiqh Books," *J. Artif. Intell. Softw. Eng.*, vol. 5, no. 2, pp. 626–635, 2025, doi: 10.30811/jaise.v5i2.7005.

[9] N. A. M. Herwanza, N. S. Harahap, F. Yanto, and F. Insani, "Penerapan Langchain Retriever dengan Model Chat Openai dalam Pengembangan Sistem Chatbot Hadis Berbasis Telegram," *J. Teknol. Inf. dan Multimed.*, vol. 6, no. 1, pp. 70–83, 2024.

[10] M. I. Syah, "Penerapan Retrieval Augemented Generation Menggunakan Langchain Dalam Pengembangan Sistem Tanya Jawab Hadis Berbasis Web," *Zo. Sist. Inf.*, vol. 6, no. 2, pp. 370–379, 2024.

[11] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "QASiNa: Religious Domain Question Answering Using Sirah Nabawiyah," *2023 10th Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2023*, 2023, doi: 10.1109/ICAICTA59291.2023.10390123.

[12] Z. Khalila *et al.*, "Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, pp. 1361–1371, 2025, doi: 10.14569/IJACSA.2025.01602134.

[13] M. A. Asl and B. Minaei-Bidgoli, "FARSIQA: Faithful & Advanced RAG System for Islamic Question Answering," 2025.

[14] A. Grattafiori *et al.*, "The llama 3 herd of models," *arXiv Prepr. arXiv2407.21783*, 2024.

[15] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models.," *Iclr*, vol. 1, no. 2, p. 3, 2022.

[16] T. J. Ahalpara, "Tell Me: An LLM-powered Mental Well-being Assistant with RAG, Synthetic Dialogue Generation, and Agentic Planning," 2025, [Online]. Available: http://arxiv.org/abs/2511.14445

[17] Y. M. Tamm, R. Damdinov, and A. Vasilev, *Quality metrics in recommender systems: Do We calculate metrics consistently?*, vol. 1, no. 1. Associationfor Computing Machinery, 2021. doi: 10.1145/3460231.3478848.

[18] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," *EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Syst. Demonstr.*, pp. 150–158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.