ZONAsi Jurnal Sistem Informasi
ZONAsi VOL. 6 NO. 2 MEI 2024
Page: 370 - 379
ISSN: 2656-7407 ( Online )
2656-7393 ( Print )

# PENERAPAN RETRIEVAL AUGEMENTED GENERATION MENGGUNAKAN LANGCHAIN DALAM PENGEMBANGAN SISTEM TANYA JAWAB HADIS BERBASIS WEB

**Muhammad Irfan Syah[1], Nazruddin Safaat Harahap[2], Novriyanto[3], Suwanto Sanjaya[4]**
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
Jl. HR. Soebrantas No 155 Km. 15, Simpang Baru, Kota Pekanbaru, Riau 28293
e-mail: [1]12050117061@students.uin-suska.ac.id , [2]nazruddin.safaat@uin-suska.ac.id , [3]novriyanto@uin-suska.ac.id , [4]suwantosanjaya@uin-suska.ac.id

***Abstrak***

*Hadis ajaran kedua setelah al-Qur'an yang menjadi panduan bagi umat Islam. Pencarian hadis saat ini kurang interaktif dalam menjawaban pertanyaan, dimana hanya menampilkan dokumen relevan. Penelitian ini bertujuan untuk mengembangkan sistem tanya jawab hadis berbasis web dengan menerapkan Retrieval Augmented Generation menggunakan framework LangChain yang diintegrasikan dengan Large Language Model GPT-4-1106-preview dari OpenAI. Sistem ini dirancang untuk membantu pengguna dalam mencari jawaban yang sesuai dengan 9 kitab hadis. Hasil penelitian menunjukkan bahwa model dapat bekerja sesuai dengan instruksi dan data dengan menyertakan sumber dari hadis terkait. Pengujian dilakukan dengan menguji 10 pertanyaan seputar hadis dengan framework BERTScore dan uji Evaluasi kualitas jawaban dengan mahasiswa ushulludin. Pada pengujian BERTScore rata-rata f1 score sebesar 0,7962, yang menunjukkan kemiripan antara jawaban sistem dengan referensi, pengujian pada Evaluasi kualitas jawaban mencapai persentase akurasi 89,4% yang menunjukkan bahwa responden "Sangat Setuju" terhadap jawaban yang dihasilkan oleh sistem.*

***Kata kunci:*** *Hadis, Langchain, Large Languange Model, Retrieval Augmented Generation, Sistem Tanya Jawab*

***Abstract***

*The second teachings after the Qur'an that guide the Muslim community are Hadiths. The current Hadith search lacks interactivity in answering questions, often only displaying relevant documents. This research aims to develop a web-based Hadith question-answering system by implementing Retrieval Augmented Generation using the LangChain framework integrated with the Large Language Model GPT-4-1106-preview from OpenAI. The system is designed to assist users in finding answers from 9 collections of Hadiths. The research results indicate that the model can work according to instructions and data by including sources from related Hadiths. Testing was conducted with 10 Hadith-related questions using the BERTScore framework, and the evaluation of answer quality was done with ushulludin students. In BERTScore testing, the average f1 score was 0.7962, indicating similarity between the system's answers and references. Evaluation of answer quality achieved an accuracy percentage of 89.4%, indicating that respondents "Strongly Agree" with the answers generated by system.*

***Keywords:*** *Hadith, Langchain, Large Language Model, Question Answering System, Retrieval Augmented Generation*

## 1. PENDAHULUAN

Dalam Islam, hadis merujuk pada segala ucapan, tindakan, atau persetujuan yang dianggap sah secara hukum oleh Nabi Muhammad SAW [1]. Sebagai sumber ajaran kedua setelah al-Qur'an, hadis menjadi panduan bagi umat Islam dalam akidah, ibadah, akhlak, dan kehidupan di dunia sebagai persiapan untuk akhirat. Pentingnya hadis sebagai tuntunan hidup didasarkan pada firman Allah dalam

QS. al-Anfal/8:20 [2]. Informasi hadis dapat diperoleh melalui konsultasi dengan ahli hadis dan pencarian buku hadis di perpustakaan, walaupun mayoritas berisi teks dalam bahasa Arab tanpa transliterasi, memerlukan pemahaman bahasa Arab sebagai keterampilan utama.

Tantangan utama dalam mengakses hadis adalah kesulitan mencari hadis yang relevan dengan topik tertentu [3]. Pencarian hadis saat ini bersifat statis dan kurang interaktif, hanya menampilkan dokumen yang menjawab pertanyaan pengguna, sehingga menyulitkan pengguna untuk mencari kembali jawaban dari pertanyaannya berdasarkan dokumen tersebut.

Terdapat konten-konten tentang perilaku beragama yang tersedia di *internet*, minat dan perhatian masyarakat dalam mencari dan mempelajari pengetahuan tentang perilaku beragama di internet masih terbatas . Situs *web* keislaman seringkali dijadikan referensi bagi umat Islam dalam mengatasi berbagai tantangan yang mereka hadapi. Penggunaan platform website tidak hanya memfasilitasi pencarian informasi tetapi juga dapat diarahkan untuk meningkatkan pemahaman terhadap agama, khususnya melalui studi hadis. *Website* dapat digunakan sebagai sarana untuk menyebarkan informasi mengenai hadis [4].

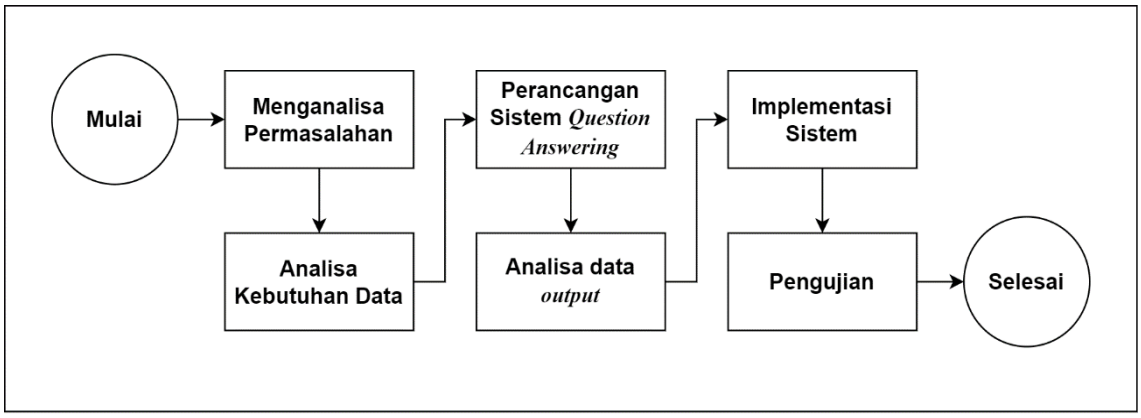
Penelitian terhadap sistem tanya jawab dilakukan oleh topsakal dan akinci [5]menggunakan *LangChain* sebagai kerangka kerja untuk mengembangkan aplikasi berbasis *Large Language Models* (LLMs) dengan menggunakan model *Chat* dari *OpenAI* dan *LangChain*. *LangChain* digunakan untuk mencari dan menampilkan sumber dari jawaban yang dihasilkan oleh sistem. Penelitian siragusa dan pirone [6] menggunakan *UnipaGPT* model *gpt-3.5-turbo* dan pendekatan *Retrieval Augmented Generation* (RAG), yang bertujuan untuk membangun *chatbot* untuk membantu siswa sekolah menengah dalam memilih program sarjana. Disisi lain levonian [7] menggunakan *OpenAI* Model *gpt-3.5-turbo-0613* dan *Retrieval Augmented Generation* (RAG) untuk membangun sistem tanya jawab matematika pada siswa sekolah menengah. Penelitian mengenai tanya jawab dengan data hadis Shahih Bukhari dilakukan oleh Asad Abdia dkk [8] yang berfokus pada metode mengekstrak Hadis yang lebih relevan untuk pertanyaan tertentu dengan *Linguistic Knowledge* (ASHLK).

Berbeda dengan penelitian sebelumnya, penelitian ini bertujuan untuk mengembangkan sebuah sistem tanya jawab hadis berbasis *web* yang memfasilitasi pengguna dalam mencari jawaban yang akurat dan relevan seputar 9 kitab hadis. Sistem ini akan mengintegrasikan *retrieval augmented generation* menggunakan *LangChain* dan model *Chat* dari *OpenAI*. Integrasi ini bertujuan untuk meningkatkan kemampuan sistem dalam menghasilkan jawaban yang informatif. Selain itu, setiap jawaban yang dihasilkan juga akan menyertakan sumber hadis untuk memastikan keakuratan dan keandalan informasi yang disediakan kepada pengguna.

Penelitian ini menggunakan 9 kitab hadis yang disebut Kutub al-Tis'ah yaitu Shahih Bukhari, Shahih Muslim, Sunan Abu Daud, Sunan Tirmidzi, Sunan Nasa'i, Sunan Ibnu Majah, Muwatha' Malik, Musnad Ahmad, dan Sunan Darimi, dengan total lebih dari 62.000 hadis [9].

## 2. METODE PENELITIAN

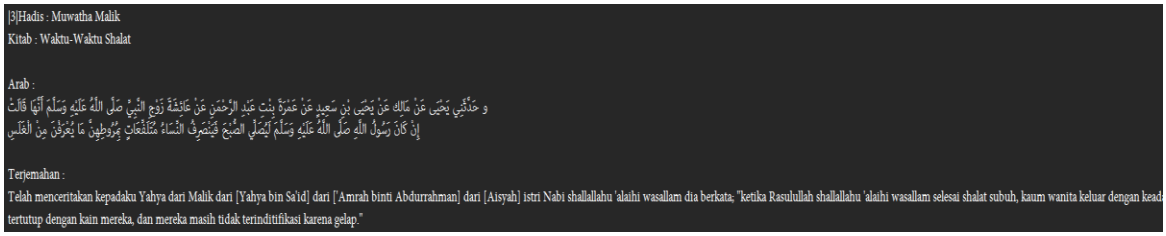
Metodologi penelitian adalah sekumpulan penjelasan dari tahapan-tahapan yang akan dilakukan dalam penelitian ini, metodologi penelitian yang akan dilakukan dapat dilihat pada gambar 1.




**Gambar 1**. Metodologi Penelitian

## 2.1. Analisa Kebutuhan Data

Analisis kebutuhan data dilakukan untuk mengevaluasi informasi yang diperlukan dalam penelitian ini. Proses pengumpulan data melibatkan pemilihan elemen data seperti nomor hadis, nama kitab, teks Arab, dan terjemahan. Dapat dilihat pada gambar 2.


[3]Hadis : Muwatha Malik
Kitab : Waktu-Waktu Shalat

Arab :
و حَدَّثَنِي يَحْيَى عَنْ مَالِك عَنْ يَحْيَى بْنِ سَعِيدٍ عَنْ عَمْرَةَ بِنْتِ عَبْدِ الرَّحْمَنِ عَنْ عَائِشَةَ زَوْجِ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ أَنَّهَا قَالَتْ
إِنْ كَانَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ لَيُصَلِّي الصُّبْحَ فَيَنْصَرِفُ النِّسَاءُ مُتَلَفِّعَاتٍ بِمُرُوطِهِنَّ مَا يُعْرَفْنَ مِنْ الْغَلَسِ

Terjemahan :
Telah menceritakan kepadaku Yahya dari Malik dari [Yahya bin Sa'id] dari ['Amrah binti Abdurrahman] dari [Aisyah] istri Nabi shallallahu 'alaihi wasallam dia berkata; "ketika Rasulullah shallallahu 'alaihi wasallam selesai shalat subuh, kaum wanita keluar dengan kead
tertutup dengan kain mereka, dan mereka masih tidak teriditifikasi karena gelap."

**Gambar 2**. Data Awal

Pada gambar di atas, semua hadis tergabung dalam satu *file* yang sama berdasarkan kitab 9 Imam Hadis. Dikarenakan data tersebut akan digunakan untuk pencarian semantik, maka data hadis perlu dibagi menjadi bagian-bagian (*chunking*) berdasarkan nomor hadis sebelum proses *embedding* dilakukan.

Pembagian data menjadi bagian-bagian (*chunking*) perlu dilakukan agar mempermudah proses pengindeksan dan pencarian dalam sistem pencarian semantik. Dengan membagi data menjadi bagian-bagian, pengguna akan dapat mengakses informasi yang relevan dengan lebih cepat dan efisien, serta meminimalkan waktu pencarian. Selain itu, pembagian data juga dapat membantu dalam mengelola sumber daya komputasi dengan lebih efektif, terutama ketika menangani volume data yang besar. Tahap selanjutnya dilakukan *embedding* oleh *OpenAI Embeddings*. *OpenAI embedding* adalah metode merepresentasikan teks dalam vektor numerik dengan makna semantik, menggunakan model bahasa yang dilatih secara luas [10], [11]. Data yang tealh di *embedding* akan disimpan didalam *database vector chroma*. *ChromaDB* adalah sebuah *database* vektor yang digunakan untuk menyimpan dokumen domain tertentu dalam bentuk *chunk-chunk* kecil yang dapat diambil sebagai konteks yang relevan dalam proses tanya jawab [12].

## 2.2 Perancangan Sistem *Question Answering*

Tahap ini melibatkan perancangan sistem yang mengintegrasikan *LangChain* dan Model *Chat OpenAI*. Pada model *Chat OpenAI*, *ChatGPT* mampu memberikan respon instan dan menyediakan informasi dengan meniru percakapan manusia [13]. GPT (*Generative Pre-trained Transformer*) adalah model bahasa yang telah dilatih sebelumnya oleh *OpenAI*, menggunakan *decoder Transformer* dengan *self-attention multi-head* dan lapisan jaringan saraf *feed-forward*. GPT memiliki kemampuan unggul dalam *pre-training generatif* [14]. GPT-3.5 dan GPT-4, *Language Models* (LMs) yang dirancang oleh *OpenAI*, menonjol dalam berbagai tugas NLP seperti terjemahan, ringkasan, dan jawaban pertanyaan, dengan dasar pada teknik deep learning yang kompleks [15]. Integrasi ini menerapkan teknik-teknik NLP (*Natural Language Processing*) yang untuk meningkatkan kualitas interaksi sistem dengan pengguna, termasuk pemahaman, pemprosessan, dan penganalisissan bahasa manusia [16]. Metode NLP ini didorong oleh pendekatan berbasis data, yang menghasilkan hasil yang lebih baik dan lebih mudah diimplementasikan [17]. Tujuan integrasi ini adalah untuk meningkatkan kinerja sistem dalam memberikan jawaban yang efektif dan menampilkan sumber jawaban dengan jelas.

*LangChain* adalah sebuah alat yang menggunakan algoritma pemrosesan bahasa alami [18], *LangChain* memudahkan pembuatan Sistem Tanya jawab dan aplikasi AI/LLM [19]. *LangChain* berfungsi sebagai kerangka kerja yang digunakan untuk mengintegrasikan RAG (*Retrieval-Augmented Generation*) dengan LLMs (*Large Language Models*). RAG bertugas mencari dokumen yang relevan terkait pertanyaan pengguna [20] sebelum diteruskan ke LLM untuk generasi jawaban[21], [22]. Dalam konteks penggunaan LLMs, terdapat perbandingan dengan model yang disesuaikan (*fine-tuned models*), yang merupakan model bahasa yang lebih kecil dan dilatih sebelumnya kemudian disesuaikan lebih lanjut pada dataset yang lebih kecil untuk meningkatkan kinerjanya pada tugas tersebut [23]. LLMs akan diintegrasikan dalam aplikasi *web* yakni Sistem Tanya-Jawab, untuk memfasilitasi interaksi pengguna dengan antarmuka bahasa alami [24]. Dengan demikian, sistem dapat memberikan jawaban yang lebih akurat dan relevan berdasarkan pertanyaan pengguna dengan baik [25].

### 2.3 Analisa data *output*

Pada tahap ini, akan dilakukan analisis terhadap jawaban atau respons yang dihasilkan oleh sistem sudah sesuai dengan apa yang tertera di dalam *database* dan sistem memberikan jawaban yang relevan. *Database* ini merupakan sumber pengetahuan yang digunakan oleh sistem tanya jawab untuk merespons pertanyaan yang diajukan. Data dalam *database* ini diperoleh dari 9 kitab imam hadis.

### 2.4 Pengujian

#### 2.4.1 *BERTScore*

*BERTScore* menghitung kemiripan dari dua kalimat sebagai jumlah dari kemiripan kosinus antara *embedding* token mereka [26]. *BERTScore* mengatasi dua kesalahan umum dalam metrik berbasis n-gram. Skor lengkap membandingkan setiap kata dalam x dengan kata dalam xˆ untuk mengukur seberapa banyak kata dalam x yang berhasil ditemukan dalam xˆ (recall), dan seberapa banyak kata dalam xˆ yang berhasil ditemukan dalam x (presisi) .

$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \quad (1)$$

Dalam rumus ini, *RBert* mengukur seberapa baik token-token dalam kalimat referensi dapat ditemukan dalam kalimat kandidat. Untuk setiap token dalam kalimat referensi, *RBert* mencari token dalam kalimat kandidat yang memiliki kesamaan kosinus tertinggi dan menghitung rata-rata dari nilai kesamaan kosinus maksimum tersebut.

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \quad (2)$$

Di sisi lain, *PBert* mengukur seberapa baik token-token dalam kalimat kandidat dapat ditemukan kembali dalam kalimat referensi. PBert melakukan proses yang serupa dengan *RBert*, tetapi dengan mencari token dalam kalimat referensi yang paling mirip dengan setiap token dalam kalimat kandidat.

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (3)$$

*FBert* menggabungkan *PBert* dan *RBert* dengan menghitung skor F1, yang merupakan *harmonic mean* dari *precision* dan *recall*. *FBert* memberikan keseimbangan antara *precision* dan *recall*, di mana nilai *FBert* yang lebih tinggi menunjukkan kesamaan yang lebih baik antara kalimat kandidat dan referensi.

#### 2.4.2 Evaluasi Kualitas Jawaban

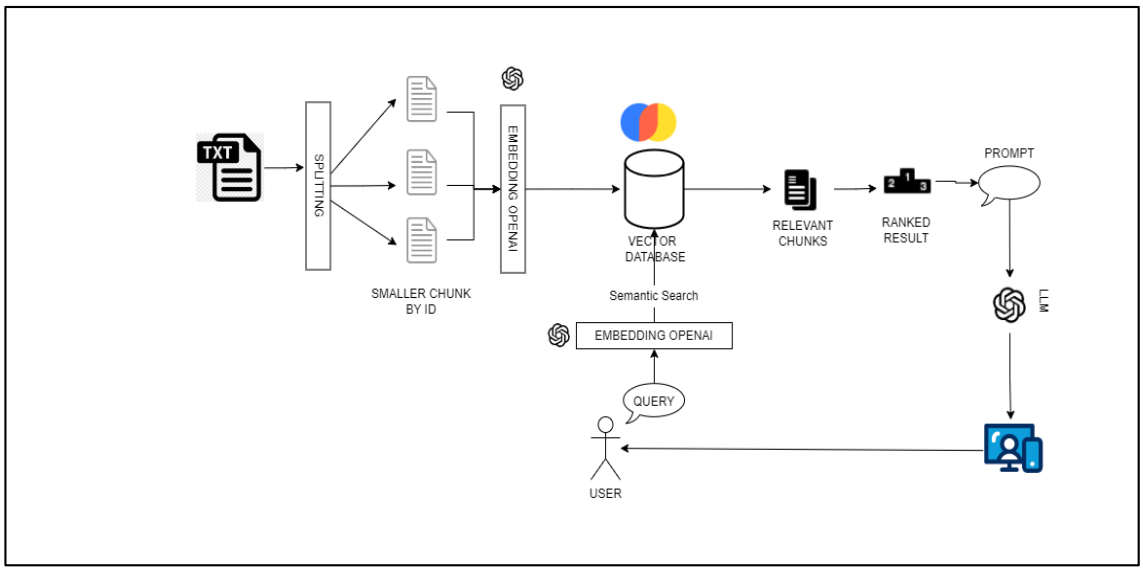
Evaluasi kualitas jawaban pada sistem RAG (*Retrieval-Augmented Generation*) hadis memerlukan partisipasi mahasiswa ilmu hadis. Mahasiswa tersebut berperan dalam menilai respons yang dihasilkan oleh sistem RAG hadis. Penilaian ini bertujuan untuk memastikan kebenaran respons sesuai dengan kaidah ilmu hadis. Pada pengujian ini menggunakan skala *likert* untuk mengukur persepsi, sikap, atau pendapat seseorang[27].

## 3. HASIL DAN PEMBAHASAN

Pada tahap ini, dilakukan dua tahapan utama, yaitu tahap implementasi sistem dan tahap pengujian.

### 3.1 Implementasi Sistem

Dalam penelitian ini, database yang digunakan terdiri dari 9 kitab hadis dengan total lebih dari 62.000 hadis. Data ini akan disematkan (embedding) untuk setiap kitab hadis secara terpisah. Sistem ini dibangun menggunakan bahasa pemrograman *Python*, serta menggunakan *library* seperti *LangChain*, *ChainLit*, *OpenAI*, *ChromaDB*, dan *TikToken*. Penjelasan mengenai alur sistem dapat dilihat pada Gambar 3.




**Gambar 3.** Alur Sistem

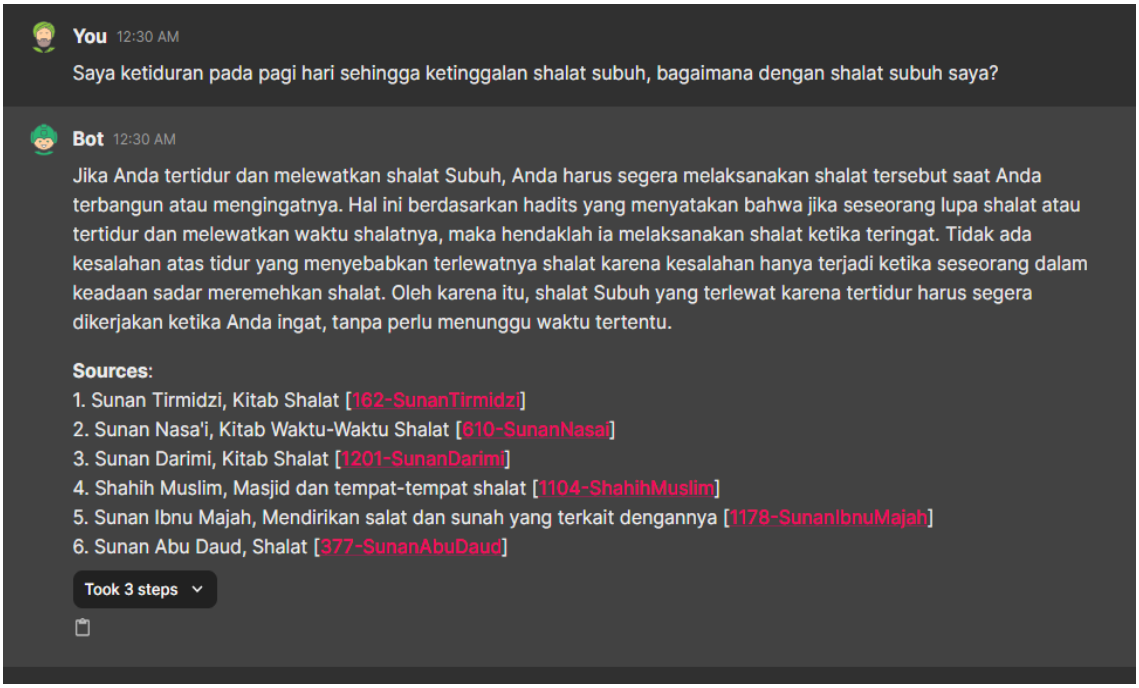
Untuk mempermudah proses pengindeksan dan pencarian dalam sistem pencarian semantik, penting untuk melakukan pembagian data menjadi bagian-bagian atau "*chunking*". Dengan membagi data menjadi bagian-bagian, pengguna dapat mengakses informasi yang relevan dengan lebih cepat dan efisien, serta meminimalkan waktu pencarian. Selain itu, pembagian data juga dapat membantu dalam pengelolaan sumber daya komputasi dengan lebih efektif, terutama ketika menangani volume data yang besar. Dalam penelitian ini, menggunakan *library re* dan fungsi *re.split* untuk memecah kumpulan *file* tersebut menjadi potongan kecil berdasarkan nomor hadis, dan kemudian menyimpan *metadata* dari setiap *chunking* tersebut untuk memungkinkan pengambilan sumber dari dokumennya. Dapat dilihat pada gambar 4.

```
[{'source': '24151-MusnadAhmad'}, {'source': '24152-MusnadAhmad'}, {'source': '24153-MusnadAhmad'}, {'source': '24154-MusnadAhmad'}]
```

**Gambar 4.** *Source Metadata*

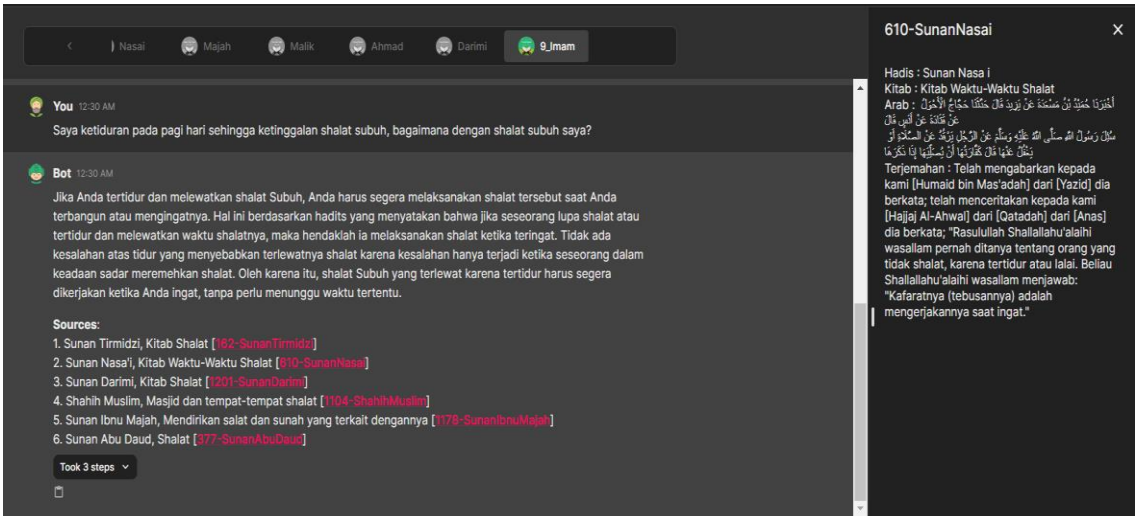
Tahapan berikutnya dalam penelitian ini melibatkan proses *encode* dan *decode* menggunakan model *Text-Embedding-3-Large*. proses *encode* merujuk pada konversi teks ke dalam representasi vektor menggunakan model *Text-Embedding-3-Large* dari OpenAI. Setiap kata dalam teks diubah menjadi vektor dalam ruang dimensi tinggi, yang mencerminkan makna semantik kata tersebut. Setelah *encode*, langkah selanjutnya adalah proses *decode*. Dalam tahap ini, vektor yang dihasilkan dari proses *encode* dikonversi kembali menjadi teks. Tujuan dari proses ini adalah untuk mempertahankan sebanyak mungkin informasi semantik dari teks asli, meskipun telah diubah menjadi representasi vektor. Kemudian hasil dari representasi vektor tersebut disimpan kedalam *database chromadb*.

Tahapan berikutnya pengguna mengajukan pertanyaan. Sistem akan melakukan pencarian semantik pada database yang telah disimpan sebelumnya. Pencarian semantik ini memungkinkan sistem untuk memahami makna di balik pertanyaan pengguna dan mencari dokumen yang paling relevan. Hasil dari pencarian semantik ini kemudian akan diurutkan, ranking teratas akan dipilih. Dengan demikian, *GPT-4-1106-preview dari OpenAI* dapat memberikan jawaban yang paling relevan kepada pengguna berdasarkan pertanyaan mereka serta menampilkan sumber dari jawaban yang diberikan. Pengguna mengajukan pertanyaan tentang hadis, kemudian sistem memberikan jawaban beserta beberapa sumber dari jawaban yang dihasilkan oleh sistem. Dapat dilihat pada gambar 5.




**Gambar 5.** Jawaban Sistem

Pada sumber yang dihasilkan, pengguna dapat mengeklik sumber dokumen yang terletak pada kurung siku. Kemudian sistem akan menampilkan content dari dokumen tersebut pada sisi kanan halaman. Dapat dilihat pada gambar 6.




**Gambar 6.** Jawaban Serta Sumber Dokumen

## 3.2 Pengujian

Pada tahap ini dilakukan dua pengujian yakni *BERTScore* dan Uji kualitas jawaban pada sistem tanya jawab:

### 3.2.1 *BERTScore*

Pada penelitian ini, menerapkan metrik evaluasi yaitu *BERTScore* untuk mencari nilai *precision*, *recall* dan f1 score antara kandidat dengan *reference*, dari total 10 pertanyaan seputar hadis, seperti yang terlihat pada tabel 1.

**Tabel 1.** Hasil Pengujian BERTScore

| No | Pertanyaan | Evaluasi | | |
|---|---|---|---|---|
| | | **Precision** | **Recall** | **F1** |
| 1 | Jika saya menemukan barang di jalan, apakah saya boleh mengambilnya? | 0.8576 | 0.8381 | 0.8477 |
| 2 | Ketika saya sedang berpuasa, saya tidak sengaja meminum air karena saya lupa, bagaimana dengan puasa saya? apakah puasa saya akan diterima? | 0.7988 | 0.7887 | 0.7937 |
| 3 | Saya ketiduran pada pagi hari sehingga ketinggalan shalat subuh, bagaimana dengan shalat subuh saya? | 0.8318 | 0.8048 | 0.8181 |
| 4 | Saat sedang berpuasa saya muntah, bagaimana dengan puasa saya? | 0.8407 | 0.8267 | 0.8336 |
| 5 | Pada saat shalat saya ingin membuang angin, namun karena sedang shalat saya menahannya, bagaimana dengan shalat saya? | 0.8275 | 0.7894 | 0.8080 |
| 6 | Saya memberi makan orang yang berpuasa, seperti apa pahala yang saya dapatkan? | 0.8496 | 0.8248 | 0.8370 |
| 7 | Ketika saya sedang shalat, kemudian ada orang yang megucapkan salam kepada saya, apa yang harus saya lakukan? | 0.8388 | 0.8363 | 0.8375 |
| 8 | Jika saya sedang melaksanakan sholat dan ada orang yang berjalan di depan saya, apa yang sebaiknya saya lakukan? | 0.8431 | 0.8187 | 0.8307 |
| 9 | Ketika saya sedang duduk, tiba-tiba ada anjing yang mendekat dan menjilati kaki saya sehingga air liur anjing tersebut mengenai kaki saya, bagaimana saya membersihkannya ? | 0.8051 | 0.7778 | 0.7912 |
| 10 | Apa Keutamaan bulan Ramadhan ? | 0.8113 | 0.8019 | 0.8066 |
| | Rata-rata | | | 0.7962 |

BERTScore mencakup nilai-nilai presisi, recall, dan skor F1. Setiap nilai ini berkaitan dengan setiap kalimat yang terdapat dalam daftar prediksi dan referensi. Rentang nilai untuk masing-masing metrik adalah dari 0.0 hingga 1.0. Presisi mengukur sejauh mana kalimat yang diprediksi relevan dengan referensi, sedangkan recall menunjukkan seberapa banyak kalimat yang relevan dari referensi berhasil diidentifikasi. Skor F1 merupakan perpaduan dari presisi dan recall, memberikan gambaran keseluruhan tentang kualitas prediksi. Pada tabel 1 diatas, Skor F1 dari masing masing pertanyaan dijadikan rata rata, dengan menjumlahkan keseluruhan Skor F1 dan kemudian membaginya dengan total pertanyaan yaitu 10. Sehingga didapatkanlah nilai rata rata dari Skor F1 sebesar 0.7962 dari 0.0 hingga 1.0.

### 3.2.2 Evaluasi Kualitas Jawaban

Dalam penelitian ini, dilakukan evaluasi kualitas jawaban terhadap pertanyaan-pertanyaan yang berkaitan dengan ajaran agama Islam. Hasil evaluasi tersebut disajikan dalam tabel 2. yang menunjukkan skor yang diberikan untuk setiap jawaban berdasarkan kriteria yang telah ditentukan. Untuk menentukan akurasi secara keseluruhan, skor yang diperoleh pada setiap kriteria diberi bobot berdasarkan tingkat kepentingannya, di mana skor SS (Sangat Setuju) dikalikan dengan 5, skor S (Setuju) dikalikan dengan 4, skor N (Netral) dikalikan dengan 3, skor TS (Tidak Setuju) dikalikan dengan 2, dan skor STS (Sangat Tidak Setuju) dikalikan dengan 1.

**Tabel 2.** Hasil Evaluasi Kualitas Jawaban

| No | Pertanyaan | Jawaban | | | | |
|---|---|---|---|---|---|---|
| | | **SS** | **S** | **N** | **TS** | STS |
| 1 | Jika saya menemukan barang di jalan, apakah saya boleh mengambilnya? | 6 | 10 | 2 | 1 | 1 |
| 2 | Ketika saya sedang berpuasa, saya tidak sengaja meminum air karena saya lupa, bagaimana dengan puasa saya? apakah puasa saya akan diterima? | 15 | 5 | 0 | 0 | 0 |
| 3 | Saya ketiduran pada pagi hari sehingga ketinggalan shalat subuh, bagaimana dengan shalat subuh saya? | 12 | 6 | 2 | 0 | 0 |
| 4 | Saat sedang berpuasa saya muntah, bagaimana dengan puasa saya? | 12 | 7 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | Pada saat shalat saya ingin membuang angin, namun karena sedang shalat saya menahannya, bagaimana dengan shalat saya? | 9 | 9 | 2 | 0 | 0 |
| 6 | Saya memberi makan orang yang berpuasa, seperti apa pahala yang saya dapatkan? | 14 | 6 | 0 | 0 | 0 |
| 7 | Ketika saya sedang shalat, kemudian ada orang yang megucapkan salam kepada saya, apa yang harus saya lakukan? | 9 | 8 | 3 | 0 | 0 |
| 8 | Jika saya sedang melaksanakan sholat dan ada orang yang berjalan di depan saya, apa yang sebaiknya saya lakukan? | 9 | 8 | 2 | 1 | 0 |
| 9 | Ketika saya sedang duduk, tiba-tiba ada anjing yang mendekat dan menjilati kaki saya sehingga air liur anjing tersebut mengenai kaki saya, bagaimana saya membersihkannya ? | 12 | 8 | 0 | 0 | 0 |
| 10 | Apa Keutamaan bulan Ramadhan ? | 15 | 5 | 0 | 0 | 0 |
| | Total | 113 | 72 | 12 | 2 | 1 |

Pada Tabel 2. diatas, setelah mengalikan skor SS, S, N, TS, dan STS dengan bobot masing-masing, total skor yang diperoleh adalah 565 untuk SS, 288 untuk S, 36 untuk N, 4 untuk TS, dan 1 untuk STS. Jumlah total dari semua skor tersebut adalah 894. Setelah itu, persentase akurasi dihitung dengan rumus total skor dibagi oleh jumlah maksimal skor yang mungkin diperoleh, lalu dikalikan dengan 100%. Dalam kasus ini, nilai x, yang didapat dari jumlah responden dikali dengan banyaknya pertanyaan dan dikali dengan skor tertinggi yang mungkin diperoleh (yaitu 5), adalah 1000.

Pada tabel 2 diatas, diperoleh persentase akurasi jawaban sebesar 89,4%. Interval penilaian yang ada pada bobot skala likert sebagai berikut :

Indeks 0% – 19,99% : Sangat Tidak Setuju
Indeks 20% – 39,99% : Tidak Setuju
Indeks 40% – 59,99% : Kurang Setuju
Indeks 60% – 79,99% : Setuju
Indeks 80% – 100% : Sangat Setuju

Nilai indeks yang didapatkan dari hasil evaluasi jawaban sebesar 89,4% maka dapat disimpulkan bahwa responden ”Sangat Setuju” terhadap jawaban yang dihasilkan oleh sistem.

### 3.3 Pembahasan

Penelitian ini menghasilkan sebuah website tanya jawab hadis dengan penerapan *retrieval augemented generation* menggunakan LangChain untuk mencari jawaban yang akurat dan relevan seputar 9 kitab hadis. Dalam meningkatkan kualitas respons, RAG digunakan sebagai pencari dokumen yang relevan menggunakan *semantic search* dalam mencari makna dari pertanyaan pengguna [21]. Penggunaan LangChain bertujuan sebagai penghubung dengan sumber data *external* yang digunakan [19], yaitu data 9 kitab hadis yang tersimpan didalam vector database ChromaDB. Langchain juga digunakan untuk mengintegrasikan sistem dengan model LLM dari *OpenAI*.

Penelitian ini berfokus untuk meningkatkan kemampuan sistem dalam menghasilkan jawaban seputar 9 kitab hadis dengan menyertakan sumber hadis yang sesuai untuk memastikan keakuratan informasi yang disediakan kepada pengguna. Sistem diuji dengan pengujian BERTScore dan uji kualitas jawaban untuk menilai kualitas arau relevansi jawaban yang dihasilkan sistem. Pada pengujian BERTScore didapatkanlah nilai rata rata dari Skor F1 sebesar 0.7962 dari 0.0 hingga 1.0. Pada uji evaluasi kualitas jawaban, peresentase indeks sebesar 89,4% mmenunjukkan responden ”Sangat Setuju” terhadap jawaban sistem.

## 4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan *Retrieval Augmented Generation* dengan menggunakan *framework LangChain* dalam pengembangan sistem tanya jawab hadis berbasis *web* telah menunjukkan hasil yang baik. Sistem ini mampu memberikan jawaban yang akurat dan relevan terhadap pertanyaan-pertanyaan seputar hadis, dengan memanfaatkan model *GPT-4-1106-preview* dari *OpenAI*. Evaluasi terhadap kualitas jawaban yang diberikan oleh sistem menunjukkan hasil yang memuaskan, dengan perolehan persentase akurasi sebesar 89,4%. Angka ini mengindikasikan bahwa responden ”Sangat Setuju” terhadap jawaban yang dihasilkan oleh sistem. Selain itu, pengujian lebih lanjut dilakukan dengan menggunakan metrik *BERTScore*, yang merupakan salah satu metode evaluasi terkini dalam mengukur kualitas teks yang dihasilkan oleh model

bahasa. Hasil pengujian menunjukkan bahwa sistem ini mampu mencapai nilai rata-rata *BERTScore* sebesar 0.7962. Pencapaian ini semakin memperkuat bukti bahwa sistem tanya jawab hadis yang dikembangkan dengan *Retrieval Augmented Generation* dan *LangChain* memiliki performa yang baik dalam menghasilkan jawaban yang berkualitas. Namun, pada penelitian berikutnya, sistem dapat ditingkatkan untuk memahami hadis dengan makna yang tidak literal dengan menyertakan penjelasan ulama, seperti tambahan dari kitab *Asbabun Wurud*. Ini akan memperluas cakupan pemahaman sistem terhadap konteks dan makna hadis, memungkinkan jawaban yang lebih tepat dan relevan..

## Daftar Pustaka


[1] M. A. Çalgan, "The Problems in adīth Usage in Kur'an Yolu Tafsīr within the Context of Qur ān-Sunnah Unity," *Cumhuriyet Ilahiyat Dergisi*, vol. 25, no. 3, pp. 1277–1298, 2021, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.18505/cuid.962041

[2] A. Supian and A. Farhan, "Pemahaman Hadis dan Implikasinya pada Praktek Keagamaan Jamaah Tabligh (Kajian Living Hadis di Kota Bengkulu)," *AL QUDS : Jurnal Studi Alquran dan Hadis*, vol. 5, no. 2, p. 537, Oct. 2021, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.29240/alquds.v5i2.2501

[3] R. Rahmatullah, "Popularitas Moderasi Beragama: Sebuah Kajian terhadap Tren Penelusuran Warganet Indonesia," *NALAR: Jurnal Peradaban dan Pemikiran Islam*, vol. 5, no. 1, pp. 62–77, Jun. 2021, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.23971/njppi.v5i1.2419

[4] R. C. Widayaningsih and M. I. Helmy, "The Fiqh Al-Hadith Of Digital Media: The Method Of Hadith Understanding Of The Website Bincangsyariah.com And Its Contribution To The Moderate Islam Discourse," *Jurnal Ushuluddin*, vol. 29, no. 2, p. 163, Dec. 2021, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.24014/jush.v29i2.13954

[5] O. Topsakal and T. C. Akinci, "Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast," *All Sciences Proceedings*, 2023.

[6] I. Siragusa and R. Pirrone, "Conditioning Chat-GPT for information retrieval: the Unipa-GPT case study," *Seventh Workshop on Natural Language for Artificial Intelligence*, 2023.

[7] Z. Levonian *et al.*, "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," *arXiv preprint*, Oct. 2023, Accessed: Nov. 08, 2023. [Online]. Available: arxiv:2310.03184v1

[8] A. Abdi, S. Hasan, M. Arshi, S. M. Shamsuddin, and N. Idris, "A question answering system in hadith using linguistic knowledge," *Comput Speech Lang*, vol. 60, Mar. 2020, doi: 10.1016/j.csl.2019.101023.

[9] A. Zubir Rosdi, S. Najihuddin Syed Hassan, N. Asiah Fasehah Muhamad, N. Izzatul Huda Mohamad Zainuzi, M. Shiham Mahfuz, and F. Pengajian Quran dan Sunnah, "PANDUAN ASAS KAEDAH KENAL PASTI STATUS HADIS: KAJIAN DISKRIPTIF PENGGUNAAN ENSIKLOPEDIA HADIS 9 IMAM (Basic Methods in Identifying the Status of Hadith: A Descriptive Overview on the Use of Encyclopedia of Hadith of the Nine Imams)," *JOURNAL OF QUR'AN AND HADITH STUDIES*, vol. 8, no. 1, pp. 2550–1488, 2023, Accessed: Dec. 25, 2023. [Online]. Available: doi:10.33102/johs.v8i1.225

[10] A. Arora and M. Dell, "LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models," *arXIv preprint*, Sep. 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2309.00789

[11] A. Kean Gao, "Vec2Vec: A Compact Neural Network Approach for Transforming Text Embeddings with High Fidelity," *ArXiv preprint*, 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2306.12689

[12] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, Accessed: Feb. 22, 2024. [Online]. Available: arXiv:2005.14165v4

[13] D. Najafali, J. M. Camacho, L. G. Galbraith, E. Reiche, A. H. Dorafshar, and S. D. Morrison, "Ask and You Shall Receive: OpenAI ChatGPT Writes Us an Editorial on Using Chatbots in Gender Affirmation Surgery and Strategies to Increase Widespread Adoption," *Aesthetic Surgery Journal*, vol. 43, no. 9. Oxford University Press, pp. NP715–NP717, Sep. 01, 2023. Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.1093/asj/sjad119

[14] S. Y. Yoo and O. R. Jeong, "EP-Bot: Empathetic Chatbot Using Auto-Growing Knowledge Graph," *Computers, Materials and Continua*, vol. 67, no. 3, pp. 2807–2817, Mar. 2021, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.32604/cmc.2021.015634

[15] D. Nunes, R. Primi, R. Pires, R. Lotufo, and R. Nogueira, "Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams," *ArXiv preprint*, Mar. 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2303.17003

[16] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural Language Processing Advancements By Deep Learning: A Survey," *arXiv preprint*, Mar. 2020, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2003.01200

[17] M. Abbaszade, V. Salari, S. S. Mousavi, M. Zomorodi, and X. Zhou, "Application of Quantum Natural Language Processing for Language Translation," *IEEE Access*, vol. 9, pp. 130434–130448, 2021, Accessed: Nov. 27, 2023. [Online]. Available: DOI:10.1109/ACCESS.2021.3108768

[18] Arjun Pesaru, Taranveer Singh Gill, and Archit Reddy Tangella, "AI assistant for document management Using Lang Chain and Pinecone," *International Research Journal of Modernization in Engineering Technology and Science*, Jun. 2023, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.56726/irjmets42630

[19] Tejaswini NR, Vidya, and Dr. T Vijaya Kumar, "LangChain-Powered Virtual Assistant for PDF Communication," *International Research Journal of Modernization in Engineering Technology and Science*, Jul. 2023, Accessed: Nov. 27, 2023. [Online]. Available: Doi:10.56726/irjmets43587

[20] Y. H. Ke *et al.*, "Development and Testing of Retrieval Augmented Generation in Large Language Models-A Case Study Report," *arXiv preprint*, 2024.

[21] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint*, Feb. 2023, Accessed: Feb. 22, 2024. [Online]. Available: arXiv:2302.13971v1

[22] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," *Adv Neural Inf Process Syst*, vol. 35, Mar. 2022, Accessed: Feb. 22, 2024. [Online]. Available: arXiv:2203.15556v1

[23] J. Yang *et al.*, "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond," *arXiv preprint*, Apr. 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2304.13712

[24] R. Pedro, D. Castro, P. Carreira, and N. Santos, "From Prompt Injections to SQL Injection Attacks: How Protected is Your LLM-Integrated Web Application?," *arXiv preprint*, Aug. 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2308.01990

[25] Y. Xie, C. Yu, T. Zhu, J. Bai, Z. Gong, and H. Soh, "Translating Natural Language to Planning Goals with Large-Language Models," *arxiv preprint*, Feb. 2023, Accessed: Nov. 27, 2023. [Online]. Available: arXiv:2302.05128

[26] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019, doi: 10.18653/V1/N19-1423.

[27] V. H. Pranatawijaya, W. Widiatry, R. Priskila, and P. B. A. A. Putra, "Penerapan Skala Likert dan Skala Dikotomi Pada Kuesioner Online," *Jurnal Sains dan Informatika*, vol. 5, no. 2, pp. 128–137, Dec. 2019, doi: 10.34128/jsi.v5i2.185.