Question Answering Al-Qur'an dan Tafsir Menggunakan Generative Pre-Trained Transformer 4 (GPT-4)

Bayu Dwinata Putra Yatabri*[1], Nazruddin Safaat Harahap[2], Lestari Handayani[3], Reski Mai Candra[4]

[1,2,3,4]Universitas Islam Negeri Sultan Syarif Kasim Riau; Jl. HR Soebrantas No. 155 Simpang Baru, Panam, Pekanbaru

[1,2,3,4]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru

e-mail: ***[1]11850112425@students.uin-suska.ac.id**, [2]nazruddin.safaat@uin-suska.ac.id, [3]lestari.handayani@uin-suska.ac.id , [4]reski.candra@uin-suska.ac.id

Abstrak

Al-Quran merupakan kitab suci yang terdiri dari 30 juz, 114 surah, 6326 ayat, dan lebih kurang 80 ribu kata Arab yang menjadi pedoman hidup bagi umat manusia. Banyaknya data dalam Al-Qur'an menjadikan tafsir sebagai pintu utama bagi umat muslim untuk memahami isi dan makna dari ayat-ayatnya. Tentu dengan ribuan kata Arab di Al-Qur'an dan penjelasan dari tafsir membuat pencarian dan pemahaman yang lebih lama, sehingga dibutuhkan sebuah sistem yang mampu memproses dan memahami data yang banyak. Penelitian ini bertujuan untuk membuat sistem yang mampu memahami dan memproses banyak data, dengan bantuan dari teknologi seperti Artificial Intelligence (AI), LangChain dan Large Language Model untuk mampu melakukan tanya jawab dalam bahasa alami. Sistem ini diuji menggunakan 2 jenis pengujian, yakni pengujian oleh user dan pengujian oleh framework DeepEval dan BertScore. Hasil dari pengujian User Acceptance Testing (UAT) berupa skor akurasi sebesar 90.28%. Dilakukan juga pengujian oleh framework DeepEval berupa hallucination dengan score sebesar 15.00%, precision sebesar 78.30% dan answer relevancy sebesar 97.20%. Selain itu terdapat pengujian menggunakan BertScore berupa score precision sebesar 75.59%, recall sebesar 62.12% dan F1 Score 72.15%.

***Kata kunci**—Al-Qur'an, Artificial Intelligence, GPT-4, Question Answering, Tafsir*

Abstract

The Quran is a holy book consisting of 30 chapters, 114 chapters, 6326 verses, and approximately 80 thousand Arabic words that guide human life. Large amount of data in the Quran makes tafsir the main door for Muslims to understand the contents and meaning of its verses. With thousands of Arabic words in the Quran and explanations of the interpretation, it takes longer to search and understand, so a system is needed to process and understand large amounts of data. This study aims to create a system that can understand and process large amounts of data, with the help of technologies such as Artificial Intelligence (AI), LangChain and Large Language Models to be able to ask questions in natural language. This system was tested using two types of testing, namely testing by users and testing by the DeepEval and BertScore frameworks. The User Acceptance Testing (UAT) test results is an accuracy score of 90.28%. Testing using DeepEval framework in the form of hallucination with a score of 15.00%, precision of 78.30% and answer relevancy of 97.20%. In addition, a test using BertScore with a precision score of 75.59%, recall of 62.12% and F1 Score of 72.15%.

***Keywords—** Al-Qur'an, Artificial Intelligence, GPT-4, Question Answering, Tafsir*

1. PENDAHULUAN

Teknologi-teknologi yang berkembang saat ini telah banyak menggunakan *Artificial Intelligence* (AI) sebagai alat dalam membantu perkejaan manusia dalam menjalankan hidupnya [1]. Bidang-bidang yang dapat ditunjang bukan hanya sebatas teknologi saja, akan tetapi sudah merambat ke berbagai bidang yang ada sesuai dengan kebutuhan manusia itu sendiri secara akurat dan cepat [2]. Ini membuat AI terus berkembang pesat dalam melakukan pemecahan masalah layaknya seperti kepintaran manusia yang dapat memecahkan rintangan di dalam kehidupan sehari-hari [3]. Dengan adanya kepintaran tersebut, AI mampu mampu menggantikan pekerjaan manusia salah satunya melakukan pencarian yang bahkan melebihi kemampuan manusia. Contohnya manusia membutuhkan waktu yang lama dalam pencarian manual terhadap data yang begitu banyak seperti Al-Qur'an dan Tafsir. Namun dengan adanya AI membuat pencarian lebih mudah dan cepat.

Al-Qur'an merupakan kitab suci yang terdiri dari 30 juz, 114 surah, 6326 ayat, dan lebih kurang 80 ribu kata Arab yang menjadi pedoman hidup umat manusia [4]. Dengan data yang begitu banyak diperlukan tafsir sebagai gerbang untuk mempelajari isi kandungan serta memahami ayat-ayat Al-Qur'an tersebut [5]. Tujuannya agar pemahaman yang didapatkan tentang makna-makna yang ada dalam kita suci tersebut dapat lebih luas dipahami berdasarkan konteks yang berkaitan seperti sejarah, linguistik, dan kulturalnya [6]. Tentu dengan data tambahan bukan hanya Al-Qur'an, melainkan beserta tafsir membuat pencarian dan pemahaman lebih lama. Maka dibutuhkan suatu sistem yang dapat dengan cepat memahaminya.

Question Answering System (QAS) menjadi salah satu solusi yang dapat digunakan untuk mengatasi masalah dalam mencari dan meminta informasi yang tepat dan relevan [7]. Ini dikarenakan dalam pencarian informasi terjadi komunikasi dua arah [8]. Selain itu penggunaan bahasa dalam QAS adalah bahasa alami dari pengguna [9]. Sehingga proses pencarian akan lebih cepat dan akurat serta lebih mudah memamami konteks yang diberikan dalam percakapan tersebut.

Penggunaan QAS biasanya dipadukan dengan *chatbot* sebagai *interface* dalam komunikasi yang dilakukan pengguna dengan sistem. *Chatbot* merupakan suatu sistem yang dapat melakukan komunikasi antara *user* dengan sistem secara langsung yang diatur oleh *robot* maupun *virtual* [10]. Implementasi *chatbot* banyak digunakan dalam penelitian, salah satunya penelitian yang berjudul Pemanfaatan *Bot Telegram* sebagai Media Informasi Akademik di STMIK Hang Tuah Pekanbaru [11]. Penggunaan *chatbot* terbukti dapat menghasilkan informasi yang dicari lebih cepat dan dapat diakses dimana saja. Sehingga penerapannya dapat diterapkan dalam sistem tanya jawab yang akan dibuat nantinya.

Penelitian yang juga menerapkan *Question Answering System* (QAS) pernah dilakukan oleh Made Bagus Putra Salabi dan kawan-kawan. Penelitian ini berbasis aturan untuk setiap pertanyaan yang diberikan terhadap data yang relevan dengan menggunakan bahasa Bali. Berbeda dengan penelitian yang dilakukan saat ini. Dengan menggunakan GPT *(Generative Pre-trained Transformer)* maka pengguna bisa bertanya tanpa adanya aturan bahasa. Hasil dari penelitian yang dilakukan memperoleh akurasi sebesar 40% kebenaran dari jawaban [12].

Penelitian lainnya terkait *Question Answering System* juga pernah dilakukan oleh Rajif Agung Yunmar dan I wayan Wiprayoga Wisesa pada tahun 2020. Adapun hasil yang didapatkan lebih bagus dari pada penelitian yang dilakukan oleh Made Bagus dan kawan-kawan. Akurasi yang dihasilkan dari penelitian tersebut yaitu sebesar 82.14% [13]. Ini dikarenakan penggunaan basis ontologi akan membuat pertanyaan yang diajukan dapat beragam dan tidak terpaku kepada aturan.

Question answering berhubungan dengan *Large Language Model* (LLM) sebagai model untuk pemrosesan. Penelitian terkait LLM pernah dilakukan oleh Waad Alshammari pada tahun 202. LLM terbukti dapat melakukan proses tanya jawab yang lebih bagus dibanding penelitian sebelumnya yang hanya menggunakan *rule-based* atau aturan. Ini dibuktikan dengan akurasi yang dihasilkan yaitu 94,45% pada model BiLSTM serta untuk AraBERTv2 dan AraBERTv0.2 sebesar 93.10% dan 93,90% [14]. Artinya model tersebut bagus dalam penerapan sistem tanya

jawab, karena penelitian ini juga menggunakan data bahasa arab yang berhubungan dengan penelitian yang akan dilakukan dengan Al-Qur'an.

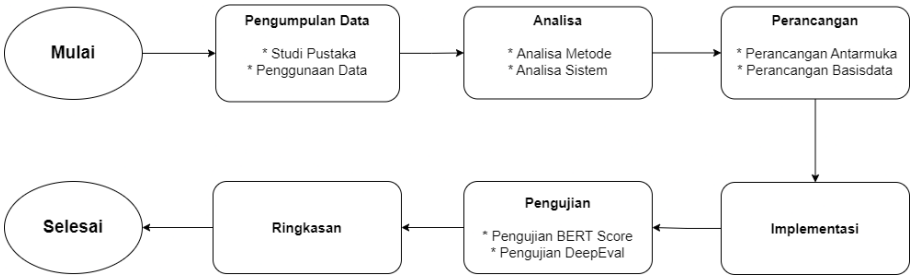
Generative Pre-Trained Transformer (GPT) sebagai model dalam memproses sistem tanya jawab yang dapat memahami bahasa secara alami khususnya terhadap Al-Qur'an dan terjemahan dalam bahasa Indonesia pernah dilakukan oleh Elvino Dwi Saputra pada tahun 2024 menggunakan GPT 3.5 sebagai model pemroses LLM. Hasil yang didapat pada penelitian tersebut 78,85% akurasi, 98,3% *answer relevancy*, dan *hallucination* sebesar 22,5% [15]. Namun perbedaan yang dilakukan pada penelitian ini yaitu menggunakan model lanjutan dari GPT yaitu dengan GPT 4.0. Sehingga diharapkan dapat menghasilkan akurasi yang lebih tinggi dibandingan GPT 3.5.

Penelitian lainnya terkait tafsir menggunakan GPT 3.5 juga pernah dilakukan oleh Febrian Rizki Adi Sutiyo pada tahun 2024. Penelitian ini menggunakan *LangChain* sebagai *framework* (kerangka kerja) agar memudahkan proses LLM. Akurasi yang didapatkan pada penelitian ini mencapai 84,29% [16]. Akurasi didapatkan dari penyebaran kuisoner untuk perbandingan jawaban GPT dengan mahasiswa jurusan Ilmu Al-Qur'an dan Tafsir UIN SUSKA. Diharapkan dengan adanya model terbaru GPT 4.0 akan membuat akurasi lebih tinggi karena model yang digunakan lebih baru dalam pemrosesan bahasa alami.

2. METODE PENELITIAN

2.1 Tahapan Review

Langkah-langkah yang akan dilakukan pada penelitian ini adalah :



Gambar 1 Tahapan Penelitian

Proses pembuatan *Question Answering* Al-Qur'an dan Tafsir melewati berbagai tahapan, mulai dari tahapan pertama yakni pengumpulan data, lalu analisa, perancangan, implementasi dan yang terakhir adalah pengujian. Setelah semua tahapan dilalui, output dari penelitian ini adalah sebuah *Question Answering System* yang berfokus pada data Al-Qur'an dan Tafsir yang menggunakan *Generative Pre-trained Transformer* 4 (GPT-4).

2.2 Pengumpulan Data

Data yang akan digunakan dalam penelitian ini berupa sebuah *vector store* yang didapatkan dari penelitian-penelitian sebelumnya, yakni :

1. *Vector store* data tafsir Al-Azhar dari penelitian yang berjudul Implementasi *Question Answering System* Tafsir Al-Azhar menggunakan *LangChain* dan *Large Language Model* berbasis Chatbot telegram [17].
2. *Vector store* data tafsir Ibnu Katsir dari penelitian yang berjudul *Question Answering System Tafseer Ibnu Katsir using Large Language Model* [18].
3. *Vector strore* data tafsir Al-Jalalain dari penelitian yang berjudul Implementasi *Question Answering* berbasis Chatbot telegram pada Tafsir Al-Jalalain menggunakan *LangChain* dan LLM [16].

4. *Vector store* data Al-Qur'an dari penelitian yang berjudul *Question Answering* Al-Qur'an menggunakan *Generative Pre-Trained Transformer* 3.5 berbasis chatbot telegram [15].

Diambil dari penelitian sebelumnya, masing-masing data awalnya adalah kumpulan *file* dengan ekstensi yang berbeda beda, untuk data Al-Qur'an data awalnya berupa file JSON (*Javascript Object Notation*), sedangkan untuk data tafsir Al-Azhar, Ibnu Katsir, dan Al-Jalalain, data awalnya berupa *file* txt, kemudian diproses dan disimpan dalam *vector stores*. Data ini nantinya akan jadikan sebuah konteks rujukan dari *Question Answering* yang akan dibuat dengan menggunakan model dari GPT terbaru, yakni GPT-4.

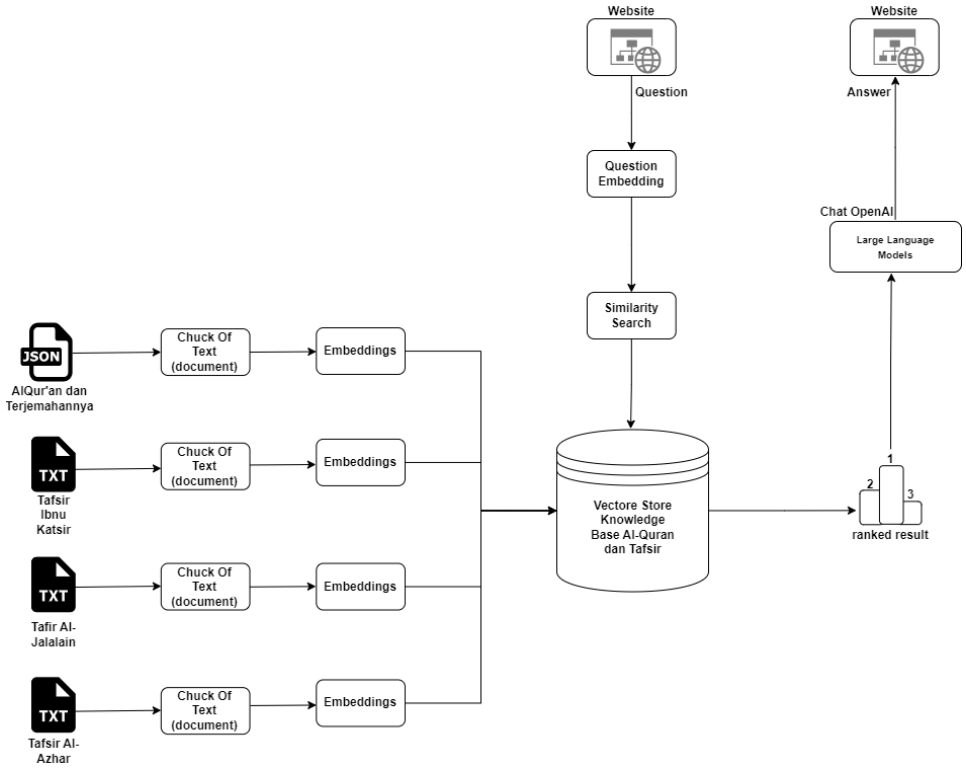
2.3 Pemrosesan Data

Pemrosesan data yang dilakukan pada penelitian ini menggunakan *framework LangChain*, ada beberapa fungsi yang disediakan oleh LangChain yang dapat digunakan untuk melakukan pemrosesan data. Tujuan pemrosesan data ini ialah agar sistem dapat mengambil ataupun melakukan pencarian pada data yang besar dengan lebih mudah [19].

a. ***Document Loader*** : komponen ini berguna untuk melakukan pengambilan dan pembacaan data. Komponen ini menyediakan beberapa fungsi yang mampu melakukan pembacaan dan pengambilan dari berbagai sumber data, seperti dari PDF, halaman *website, Plain Text* dan beberapa ekstensi file lainnya.
b. ***Text Splitter*** : *document* yang telah berhasil dimuat menggunakan fungsi dari *document loader* umumnya memiliki teks atau kalimat yang panjang sehingga tidak efektif untuk dilanjutkan pemrosesan datanya, oleh karena itu *Text Splitter* berfungsi untuk memotong dan membagi teks yang panjang tersebut menjadi beberapa potongan kecil agar lebih mudah dan lebih efektif jika diproses lebih lanjut.
c. ***Embedding Models*** : komponen ini bertujuan untuk mengkonversi teks yang telah dipotong kedalam bentuk numerik*(vector embedding)*.
d. ***Vector Stores*** : sebuah basis data yang digunakan untuk menyimpan *vector embedding* (hasil dari *embedding models*), selain menyimpan *vector stores* juga dapat membuat pencarian yang efisien berdasarkan jarak antar *vector*
e. ***Retriever*** : komponen yang bertugas mengambil data atau informasi yang relevan dari dalam *vector store*, berdasarkan pertanyaan yang di-*input*kan pengguna.

2.4 Implementasi Sistem

Setelah tahap pemrosesan data, proses selanjutnya adalah implementasi sistem. Sistem akan dibangun menggunakan bahasa pemrograman python, dan *framework* LangChain. Adapun diagram alir yang digunakan dalam penelitian ini :



Gambar 2 Diagram Alir *Question Answering System*

Data *vector store* yang ada pada proses diagram alir diatas, seperti data tafsir Al-Azhar, Al-Jalalain, Ibnu Katsir dan Al-Quran beserta terjemahan didapatkan dari beberapa penelitian sebelumnya. Adapun data yang didapatkan dalam bentuk *vector store* menggunakan database FAISS dan juga ChromaDB.

2.5 Generative Pre-Trained Transformer 4 (GPT-4)

Generative Pre-Trained Transformer 4 atau disingkat GPT-4 merupakan pengganti dari GPT-3.5 yang dirilis oleh perusahaan bernama OpenAI pada tanggal 14 maret 2023. GPT-4 menggunakan pendekatan berupa pembelajaran mesin yang kuat, agar mampu menghasilkan teks yang lebih mudah untuk dipahami dan kontekstual bagi pengguna. Model yang digunakan pada GPT-4 telah didesain dengan jutaan parameter yang telah dilatih menggunakan data dalam jumlah yang besar, sehingga mampu menghasilkan jawaban yang lebih baik [20].

GPT-4 merupakan *upgrade* dari GPT-3.5, terdapat beberapa perbedaan dari ke-2 model tersebut, seperti dari ukuran dan kompleksitas model, GPT-4 mempunyai ukuran yang lebih besar dan juga kompleks jika dibandingkan dengan GPT-3.5. Ukuran yang besar ini didapati dari banyaknya *dataset* yang telah digunakan untuk melatih model dari GPT-4 ini , sehingga, tidak hanya mampu memahami dan menghasilkan teks lebih baik dari GPT-3.5, GPT-4 mampu melakukan penyempurnaan tata bahasa menjadi lebih halus, atau lebih mudah dipahami oleh pengguna.

Penelitian ini akan menggunakan GPT-4 sebagai inti dari pemrosesan bahasa alami, gpt-4 akan mencoba memahami dan juga merespons pertanyaan yang diberikan oleh pengguna secara mendalam. GPT-4 akan berintegrasi dengan *framework* LangChain , setelah tahapan pemrosesan data selesai. GPT-4 akan digunakan untuk memproses *query* atau pertanyaan dari pengguna, lalu membantu memilih teks mana yang paling relevan untuk dijadikan sebuah jawaban. Selain membantu menghasilkan jawaban GPT-4 juga akan merangkai jawaban yang dibentuk menjadi lebih halus, dengan gaya bahasa yang mudah dipahami oleh pengguna. Dengan begitu penggunaan GPT dapat dinilai sebagai sistem yang dengan cepat dan efisien dalam pemberian dan peningkatan produktivitas suatu informasi yang dicari [21].

2.6 Pengujian Sistem

Pengujian sistem adalah tahapan yang digunakan untuk melakukan pengujian pada sistem, pada penelitian ini pengujian yang digunakan adalah pengujian sistem oleh *user*, dan yang kedua pengujian sistem menggunakan *framework*. Pengujian oleh *user* dilakukan dengan cara memberikan akses kepada *user* agar bisa melakukan interaksi langsung dengan sistem, dan nantinya *user* akan menilai sistem yang telah dibuat apakah sudah sesuai atau belum, dengan cara berinteraksi langsung dan melakukan tanya jawab dengan sistem. Adapun yang akan menjadi *user* untuk pengujian ini adalah ustadz, hal ini dilakukan agar penilaian lebih akurat karena langsung dari pakar agama.

Pengujian menggunakan *framework* yang akan digunakan pada penelitian ini, yang pertama akan dilakukan pengujian menggunakan *framework* DeepEval, dengan tujuan untuk mengetahui nilai dari *hallucination*, *answer relevancy* dan *precision* [17]. DeepEval merupakan *open source framework* yang digunakan untuk evaluasi penggunaan *Large Language Model* yang menggunakan bahasa pemrograman python [22].

a. *Hallucination* : jawaban yang dihasilkan seolah-olah relevan namun pada dasarnya salah dan tidak berasal dari data yang diberikan, atau bisa disebut model "menghalusinasikan" jawaban dan menciptakan fakta tidak akurat.
b. *Answer Relevancy* : mencari seberapa relevan jawaban dengan pertanyaan yang diajukan oleh pengguna dari topik yang diminta.
c. *Precision* : membantu menentukan seberapa tepat jawaban yang telah diberikan oleh model, dimana jika *precision* bernilai tinggi maka jawaban yang diberikan adalah benar.

Pengujian menggunakan *framework* kedua yang akan dilakukan ialah pengujian menggunakan *Bert Score*, dengan tujuan mendapatkan nilai dari *precision*, *recall* dan *F1 score*.

a. *Precision* : mengukur ketepatan model dalam menghasilkan jawaban benar dengan membandingkan seluruh jawaban yang dihasilkan.
b. *Recall* : mengukur kemampuan model dalam menemukan semua jawaban benar dari data yang tersedia.
c. *F1 score* : metrik yang tercipta dari gabungan antara *precision* dan juga *recall*, dengan tujuan untuk memastikan kalau sistem tidak hanya akurat, akan tetapi semua jawaban yang dihasilkan juga relevan.

3. HASIL DAN PEMBAHASAN

3.1 Pemrosesan Data

Data tafsir dan Al-Quran yang digunakan dalam penelitian didapatkan dari penelitian yang telah dilakukan sebelumnya. Adapun data yang digunakan berasal dari pemrosesan menggunakan *LangChain*, akan tetapi jenis penyimpanan yang digunakan berbeda. Seperti data Tafsir Al-Azhar dan Al-Quran yang menggunakan *vector store* berupa *FAISS*, lalu Tafsir Al-Jalalain dan Ibn-Kathir menggunakan vector store berupa *ChromaDB*.

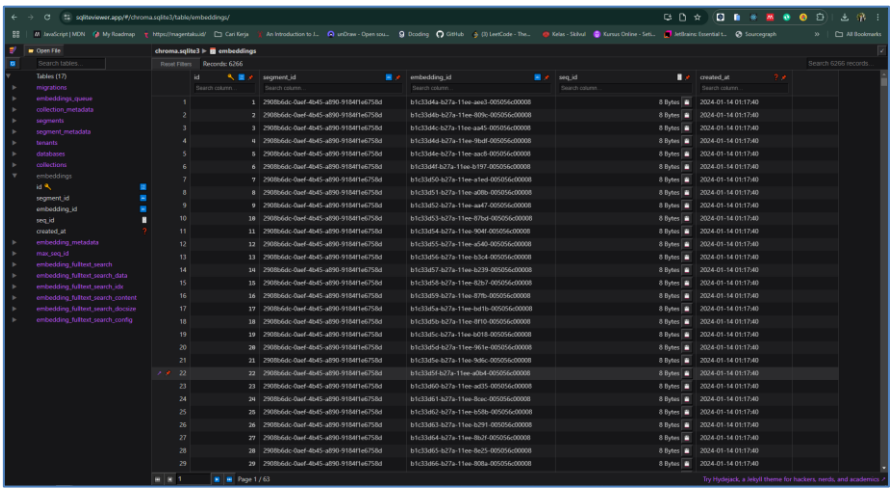
FAISS merupakan salah satu *library* dari bahasa pemrograman python yang dapat digunakan untuk melakukan penyimpanan *vector store*, begitu juga dengan *ChromaDB*. Selain digunakan sebagai sebuah penyimpanan, *FAISS* dan *ChromaDB* juga berfungsi sebagai mesin pencari data yang telah disimpan didalamnya. Dengan menggunakan *function* atau *method* seperti *similarity search*, *FAISS* dan *ChromaDB* dapat mencari data sesuai dengan kemiripan antara *vector* pertanyaan atau data yang mau dicari dengan *vector* yang telah disimpan ke dalam *FAISS* ataupun *ChromaDB*.

Data yang digunakan sebelumnya telah diproses lebih lanjut, seperti dilakukan pembersihan data, dimana dilakukan pembersihan berupa pembuangan karakter-karakter yang tidak penting. Lalu, dilakukan juga proses pemotongan teks atau *chunk of text* pada data tafsir dan

Al-Quran dan hasil dari pemotongan tersebutl di-*embed* lalu di simpan di *vector store*. Adapun isi dari data yang menggunakan penyimpanan *vector store FAISS* adalah sebagai berikut :

Gambar 3 Visual Data *FAISS*

Hasil pengolahan pada data yang telah dimasukkan dalam bentuk *vector store* didalam *database FAISS* dapat dilihat pada Gambar 3. Dimana data yang telah di-*embed* akan menjadi kumpulan *float-point* yang mewakili data aslinya. Visualisasi data tentu tidak bisa dilihat layaknya pada data asli sebelum dilakukan *embedding* data. Sedangkan data yang menggunakan penyimpanan vector store *Chromadb*, datanya dapat dilihat menggunakan tool bantuan seperti sqliteviewer yang bisa dicari di *Google*. Adapun bentuk dari chromadb yang dibuka menggunakan tool sqliteviewer dapat dilihat pada Gambar 4 dibawah ini.



Gambar 4 Visual Data *Chroma DB*

Data yang disimpan dengan *ChromaDB* biasanya akan bertipe sqlite. Sehingga jika dibuka dengan *tools* seperti diatas, akan terlihat bahwa data telah diubah menjadi *database* lengkap dengan beberapa tabel sebagai penunjang dan mewakili dari data asli yang telah diolah. Hal ini disebabkan proses *embeding* dengan penyimpanan *Chroma* berbeda dengan FAISS tapi sama-sama disimpan dalam *vector stores* yang didalamnya berupa *floating point* data aslinya.

3.2 Hasil Sistem

Hasil implementasi *Question Answering System* pada data Tafsir dan Al-Quran menggunakan GPT (*Generative Pre-Trained Transformer*) 4 menggunakan streamlit adalah sebagai berikut :

1. Tampilan user bertanya dan sistem menjawab

Gambar 5 Tampilan *Question Answering System* Al-Quran dan Tafsir

Penggunaan GPT-4 sebagai model membuat hasil yang didapat dari 4 konteks yang disediakan menjadi lebih baik dibanding menggunakan GPT-3.5, hal tersebut disebabkan oleh kemampuan GPT-4 dalam memahami bahasa alami atau pertanyaan yang ditanyakan serta konteks yang diberikan dari ke 4 *vector store* dengan baik. Menggunakan GPT-4 juga membuat *output* atau jawaban yang lebih natural sehingga jawaban yang diberikan sistem lebih dapat dimengerti oleh pengguna.

3.3 Pengujian

Pada penelitian ini dilakukan beberapa pengujian, baik dari pengujian sistem oleh user ataupun pengujian terhadap hasil jawaban *Question Answering System* menggunakan GPT-4. Ada 10 pertanyaan yang akan digunakan dalam pengujian sistem oleh 7 ustadz dan juga *framework*, pertanyaannya yakni :

Tabel 1 Pertanyaan pengujian

No	Pertanyaan
1	Bagaimana hubungan kita dengan anak yatim ?
2	Bagaimana sikap seorang muslim kepada kedua orang tua ?
3	Mengapa zakat menjadi salah satu kewajiban bagi umat islam ?
4	Apa yang dimaksud dengan ayat mutasyabihat dalam Al-Qur’an ?
5	Bagaimana Al-Qur’an menjelaskan proses penciptaan manusia dalam surah Al-Mu’minun ?
6	Bagaimana Al-Qur’an menggambarkan kehidupan di surga seperti yang di sebutkan dalam surah Ar-Rahman ?
7	Mengapa kisah Nabi Musa AS sering disebutkan berulang kali dalam Al-Quran ?

No	Pertanyaan
8	Kapan orang-orang yang beriman akan mendapatkan kemenangan, seperti yang disebutkan dalam surah An-Nur ?
9	Siapa yang bertanggung jawab menyampaikan wahyu dari Allah kepada para nabi, seperti yang disebukan dalam surah An-Najm ?
10	Apa tujuan Allah menciptakan alam semesta seperti yang di jelaskan alam Al-Qur'an ?

Dan didapat hasil perhitungan dari 10 pertanyaan, yakni :

1 = 0 x 1 = 0
2 = 0 x 2 = 0
3 = 4 x 3 = 12
4 = 26 x 4 = 104
5 = 40 x 5 = 200

Total = 316

$$x = skor\ tertinggi\ \times\ (jumlah\ pertanyaan\ \times\ jumlah\ responden) \quad (1)$$
$$x = 5\ \times (10\ \times 7)$$
$$x = 350$$

$$akurasi = \frac{total}{x} \times 100\ \% \quad (2)$$

$$akurasi = \frac{316}{350} \times 100\ \%$$

$$akurasi = \ 90{,}28\ \%$$

Dari hasil pengujian sistem yang dilakukan oleh user yakni ustadz, didapatkan nilai akurasi jawaban yang diberikan oleh sistem sebesar 90,28 %.

Pengujian selanjunya adalah pengujian terhadap sistem menggunakan *DeepEval*, dengan tujuan mencari seberapa tinggi halusinasi jawaban yang diberikan oleh sistem terhadap 4 konteks yang disajikan menggunakan GPT-4, pengujian ini juga mencari seberapa akurat jawaban yang di hasilkan oleh sistem dan seberapa relevan dengan 4 konteks yang diberikan, dan diproses menggunakan GPT-4. Berikut adalah hasil pengujian sistem menggunakan *DeepEval* :

Tabel 2 Hasil Pengujian *DeepEval*

No	Hallucination	Answer Relevancy	Precision
1	0,00 %	89, 00 %	100, 00 %
2	25,00 %	100, 00 %	50, 00 %
3	0, 00 %	83, 00 %	50, 00 %
4	25, 00 %	100, 00 %	76, 00 %
5	0, 00 %	100, 00 %	100, 00 %
6	75, 00 %	100, 00 %	93, 00 %
7	0, 00 %	100, 00 %	100, 00 %
8	0, 00 %	100, 00 %	81, 00 %
9	25, 00 %	100, 00 %	33, 00 %
10	0, 00 %	100, 00 %	100, 00 %
Rata-rata	15, 00 %	97, 20%	78, 30 %

Dari pengujian menggunakan *DeepEval*, sistem menghasilkan tingkat *hallucination* yang rendah yakni sebesar 15,00%, ini membuktikan bahwa model telah mampu menghasilkan jawaban yang sesuai dengan konteks yang diberikan, lalu skor *answer relevancy* yang tinggi sebesar 97,20% menunjukkan bahwa jawaban yang diberikan oleh sistem relevan dengan pertanyaan yang diberikan user, dan yang terakhir skor *precision* sebesar 78,30 % menunjukkan bahwa sistem cukup efektif dalam memberikan jawaban sesuai dengan pertanyaan dan konteks yang diinginkan pengguna.

Pengujian terakhir adalah pengujian terhadap model menggunakan *Bert Score*, pengujian ini digunakan untuk melakukan pengecekan apakah model mampu mengambil atau me-*retrieve* jawaban berdasarkan 4 konteks data yang diberikan, dan model yang digunakan dalam penelitian ini adalah model dari OpenAi yakni GPT-4.

Tabel 3 Hasil Pengujian *Bert Score*

No	Precission	Recall	F1 Score
1	70, 34 %	66, 70 %	68, 47 %
2	70, 86 %	63.23 %	63, 83 %
3	73, 05 %	64.31 %	68, 40 %
4	78, 08 %	72, 88 %	75, 39 %
5	74, 64 %	66, 00 %	70, 06 %
6	78, 90 %	67, 41 %	72, 71 %
7	76, 63 %	70, 89 %	73, 65 %
8	73,71 %	72, 30 %	72, 95 %
9	84,60 %	84,90 %	84,75 %
10	75,20 %	62,53 %	68,28 %
Rata-rata	75,59 %	69, 12 %	72, 15 %

Tabel diatas menunjukkan hasil pengujian model menggunakan *BertScore* terhadap model GPT-4 mendapatkan hasil rata-rata *precision* sebesar 75.59%, lalu nilai rata-rata *recall* 62.12% dan *F1 Score* 72.15%. Ini menunjukkan bahwa jawaban yang diberikan oleh sistem cenderung relevan dengan 4 konteks yang diberikan, akan tetapi masih ada beberapa jawaban yang belum sepenuhnya relevan dengan konteks yang diberikan.

Penelitian ini telah selesai dilakukan sesuai dengan tahapan penelitian yang telah direncanakan dan berhasil menunjukkan bahwa model GPT-4 yang digunakan memiliki relevansi yang cukup baik. Walaupun terdapat beberapa perbaikan yang bisa dilakukan lebih lanjut guna meningkatkan akurasi dan relevansi jawaban yang diberikan oleh sistem.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, disimpulkan bahwa penulis telah sukses dan berhasil membangun sebuah sistem yang dapat melakukan tanya jawab terhadap 4 konteks data sekaligus. Dilakukan beberapa pengujian terhadap sistem, yakni pengujian sistem oleh *user* dan pengujian sistem menggunakan framework. Pengujian sistem oleh user mendapatkan score 90,28%, hasil ini menunjukkan bahwa sistem telah memberikan jawaban yang baik serta relevan.

Pengujian selanjutnya menggunakan *DeepEval* menghasilkan nilai *hallucination* yang rendah, yaitu 15,00%. Menunjukkan bahwa model mampu menghasilkan jawaban yang sesuai dengan konteks yang telah berikan. Lalu, nilai *answer relevancy* yang tinggi sebesar 97,20% menunjukkan bahwa jawaban yang diberikan oleh sistem relevan dengan pertanyaan yang diberikan user. Dan yang terakhir nilai *precision* sebesar 78,30% yang menunjukkan bahwa

sistem cukup efektif dalam memberikan jawaban sesuai dengan pertanyaan dan konteks yang diinginkan pengguna.

Pengujian terhadap model menggunakan *BertScore* mendapatkan hasil skor berupa nilai *precision* sebesar 75.59 %, *recall* 69.12% dan nilai *F1 Score* sebesar 72.15 %. Dari kedua pengujian sistem menggunakan *framework*, didapatkan bahwa sistem telah mampu memberikan jawaban yang relevan, dan hampir di tiap jawabannya diambil dari 4 konteks data yang berikan.

5. SARAN

Penelitian selanjutnya dapat memperluas cakupan data, mungkin tidak hanya terpaku dari data Tafsir dan juga Al-Qur'an, mungkin bisa memperluas ke data-data yang lain seperti tafsir dan lain sebagainya agar sistem lebih mempunyai banyak data eksternal atau konteks, dan mampu menjawab pertanyaan-pertanyaan tidak hanya dari Al-Qur'an dan Tafsir saja, serta dapat menghasilkan jawaban dengan cepat, dan akurat.

DAFTAR PUSTAKA

[1] R. J. Sahputra and A. Muzakir, "Penerapan AI Melalui Pendekatan Heuristik Semilaritas Pada Game Edukasi Anak Usia Dini," *J. Pengemb. Sist. Inf. dan Inform.*, vol. 1, no. 4, pp. 209–219, 2021, doi: 10.47747/jpsii.v1i4.547.

[2] M. Sidik, B. Gunawan, and D. Anggraini, "Pembuatan Aplikasi Chatbot Kolektor dengan Metode Extreme Programming dan Strategi Forward Chaining," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 2, pp. 293–302, 2021, doi: 10.25126/jtiik.2021824298.

[3] T. N. Fitria, "Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay," *ELT Forum J. English Lang. Teach.*, vol. 12, no. 1, pp. 44–58, 2023, doi: 10.15294/elt.v12i1.64069.

[4] R. Malhas, W. Mansour, and T. Elsayed, "Qur ' an QA 2022 : Overview of The First Shared Task on Question Answering over the Holy Qur ' an," no. June, pp. 79–87, 2022.

[5] A. Latif, "Spektrum Historis Tafsir Al-Qur'an Di Indonesia," *TAJDID J. Ilmu Ushuluddin*, vol. 18, no. 1, pp. 105–124, 2020, doi: 10.30631/tjd.v18i1.97.

[6] F. Fatmawati, "Al-Tadabbur: Jurnal Kajian Sosial, Peradaban dan Agama STUDI PENELITIAN TAFSIR DI INDONESIA (Pemetaan Karya Tafsir Indonesia Periode 2011-2018)," pp. 81–102, 2020.

[7] L. A. Permana, N. Safaat Harahap, L. Handayani, and M. Affandes, "Implementation of Question Answering Feature in Mobile-Based Qur'an Application Using Flutter Framework," *J. Penelit. Ilmu dan Teknol. Komput.*, vol. 16, no. 1, pp. 301–312, 2020.

[8] D. Apriliani, S. F. Handayani, T. N. Anugrahaeni, A. Miftahudin, L. Nurarifiah, and I. T. Saputra, "Aplikasi Question Answer Sebagai Media Pembelajaran Interaktif Untuk Mata Pelajaran Akuntansi," *JMM (Jurnal Masy. Mandiri)*, vol. 7, no. 2, pp. 2003–2011, 2023, doi: 10.31764/jmm.v7i2.13867.

[9] A. Dhandapani and V. Vadivel, "Question Answering System over Semantic Web," *IEEE Access*, vol. 9, pp. 46900–46910, 2021, doi: 10.1109/ACCESS.2021.3067942.

[10] D. L. Dadang Iskandar Mulyana, "Implementasi Chatbot Telegram Dalam Meningkatkan Partisipasi Kegiatan Warga," *J. Pengabdi. Kpd. Masy. Nusant.*, vol. 4, no. 2, pp. 866–874, 2023.

[11] G. C. Lenardo, Herianto, and Y. Irawan, "Pemanfaatan Bot Telegram sebagai Media Informasi Akademik di STMIK Hang Tuah Pekanbaru," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 1, no. 4, pp. 351–357, 2020, doi: 10.35746/jtim.v1i4.59.

[12] M. A. P. Subali and P. Wijaya, "Sistem Question Answering untuk Bahasa Bali menggunakan Metode Rule-Based dan String Similarity," *Techno.Com*, vol. 20, no. 2, pp. 300–308, 2021, doi: 10.33633/tc.v20i2.4390.

[13] R. A. Yunmar and I. W. W. Wisesa, "Pengembangan Mobile-Based Question Answering

System Mobile-Based Question Answering System Development With Ontology Based Knowledge," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 4, pp. 693–700, 2020, doi: 10.25126/jtiik.202072255.

[14] W. Alshammari and S. Alhumoud, "TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT with BiLSTM," *IEEE Access*, vol. 10, no. August, pp. 91509–91523, 2022, doi: 10.1109/ACCESS.2022.3198955.

[15] N. S. H. J. Y. Elvino Dwi Saputra, "Question Answering Al-Qur'an Menggunakan GPT 3,5 Chatbot Telegram," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, no. No.1, pp. 550–563, 2024.

[16] F. Rizki, A. Sutiyo, N. S. Harahap, S. Agustian, and R. M. Candra, "Implementasi Question Answering Berbasis Chatbot Telegram Pada Tafsir Al-Jalalain Menggunakan LangChain dan LLM," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 5, pp. 2464–2472, 2024, doi: 10.30865/klik.v4i5.1784.

[17] Aji Bayu Permadi, Nazruddin Safaat H, Lestari Handayani, and Yusra, "Implementasi Question Answering System Tafsir Al-Azhar Menggunakan LangChain Dan Large Language Model Berbasis Chatbot Telegram," *J. Teknoif Tek. Inform. Inst. Teknol. Padang*, vol. 12, no. 1, pp. 62–69, 2024, doi: 10.21063/jtif.2024.v12.1.62-69.

[18] A. S. Prihatinoto, N. S. Harahap, M. Irsyad, and I. Iskandar, "QUESTION ANSWERING SYSTEM TAFSEER IBNU KATSIR USING LARGE LANGUAGE MODELS," *Jire J. Inform. Rekayasa Elektron.*, vol. 7, no. 1, 2024, doi: 10.36595/jire.v7i1.1155.

[19] O. Topsakal and T. C. Akinci, "Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast," *Int. Conf. Appl. Eng. Nat. Sci.*, vol. 1, no. 1, pp. 1050–1056, 2023, doi: 10.59287/icaens.1127.

[20] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A Survey on Assessments of ChatGPT and GPT-4," *2024 Jt. Int. Conf. Comput. Linguist. Lang. Resour. Eval. Lr. 2024 - Main Conf. Proc.*, pp. 7844–7866, 2024.

[21] W. Suharmawan, "Pemanfaatan Chat GPT Dalam Dunia Pendidikan," *Educ. J. J. Educ. Res. Dev.*, vol. 7, no. 2, pp. 158–166, 2023, doi: 10.31537/ej.v7i2.1248.

[22] "DeepEval - The Open-Source LLM Evaluation Framework." Accessed: Mar. 12, 2024. [Online]. Available: https://docs.confident-ai.com/