Implementasi *Transformer* dan *RAG* untuk Otomatisasi Akuisisi Informasi dari Publikasi Akademik: Konstruksi *Chatbot* Berbasis *Web* Berorientasi *PDF*

Ruth Johana Hutagalung, Totok Mujiono dan Djoko Purwanto
Departemen Teknik Elektro, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: totok_m@ee.its.ac.id

***Abstrak*—Dalam konteks tantangan berkelanjutan yang dihadapi oleh komunitas akademis dalam menyerap serta menginterpretasi konten publikasi ilmiah yang kompleks dan luas, yang memerlukan tingkat pemahaman konseptual yang tinggi dan keterlibatan kognitif yang intens, integrasi paradigma pembelajaran mesin dengan spesialisasi *Transformer* ke dalam sistem otomatisasi menjanjikan peningkatan signifikan dalam efisiensi akuisisi informasi ilmiah. *LLM* menunjukkan kapabilitas yang mengesankan tetapi dihadapkan pada kendala seperti halusinasi, pengetahuan yang usang, serta proses penalaran yang tidak transparan dan sulit dilacak. *RAG* muncul sebagai solusi potensial dengan menggabungkan pengetahuan dari *database* eksternal. Paper ini mengimplementasikan dan mengintegrasikan *OpenAI GPT-4* dan *Pinecone* ke dalam *framework NextJs* untuk konstruksi sebuah *website chatbot* yang berorientasi *PDF*. Analisis kuantitatif mengungkapkan variasi kinerja sistem dalam beberapa dimensi evaluasi. Tingkat *context precision* mencapai 0.79, menandakan efektivitas yang baik, sementara *answer relevancy* menunjukkan performa unggul dengan nilai 0.83. Kemampuan *context recall* dan *answer correctnepenss* menampilkan hasil yang cukup memuaskan, masing-masing dengan skor 0.68 dan 0.7. Meskipun demikian, aspek *faithfulness* memperoleh skor 0.51, mengindikasikan kebutuhan akan penyempurnaan signifikan dalam area ini untuk meningkatkan kinerja sistem secara keseluruhan. Integrasi berkas *PDF* pada sistem mengakibatkan peningkatan signifikan dari tingkat keakuratan jawaban sebesar 0.2, dari 0.5 pada tahap pre-implementasi menjadi 0.7 pada tahap post-implementasi.**

***Kata Kunci*—Transformer, RAG, Chatbot, Website.**

I. PENDAHULUAN

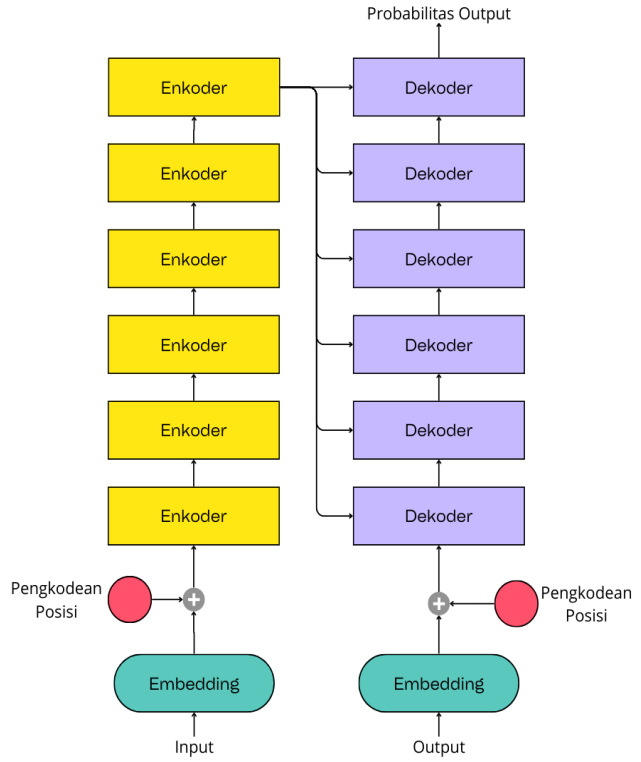
BERDASARKAN data tahun 2022, terdapat estimasi publikasi lebih dari 5,14 juta artikel ilmiah setiap tahunnya. Peneliti dan komunitas akademis secara luas menghadapi tantangan substantif dan berkelanjutan dalam asimilasi dan interpretasi kandungan publikasi ilmiah yang kompleks dan ekstensif, yang menuntut tingkat pemahaman konseptual tinggi serta keterlibatan kognitif yang intensif. Variabilitas dalam interpretasi manusia seringkali dikondisikan oleh variabel subjektif seperti kelelahan dan derajat konsentrasi, potensial untuk inkonsistensi dalam pengolahan informasi. Selanjutnya, elaborasi informasi merupakan awal dari sebuah proses analitis yang komprehensif, dimana relevansi terhadap area studi spesifik harus ditentukan melalui evaluasi yang kritis dan waktu yang signifikan untuk kontemplasi mendalam.

Dalam wacana ilmiah global, kemahiran dalam bahasa Inggris, yang secara luas diakui sebagai *lingua franca* internasional, juga menjadi kendala utama bagi para ilmuwan yang tidak memiliki bahasa Inggris sebagai keahlian linguistik primer mereka. Hingga tahun 2020, terdapat 46.736 jurnal akademik yang menerbitkan artikel ilmiah di seluruh dunia. Di tahun yang sama, terdapat 35.070 jurnal berbahasa Inggris yang diterbitkan, mewakili 75,04% dari keseluruhan jurnal akademik di dunia. Data ini menunjukkan prevalensi bahasa Inggris yang signifikan dalam publikasi ilmiah. Hambatan linguistik secara signifikan membatasi kemampuan mereka untuk mensintesis pengetahuan secara efektif, menimbulkan hambatan kritis untuk partisipasi penuh dan kontribusi dalam komunitas akademis internasional.

Integrasi paradigma *machine learning* dengan spesialisasi *Transformer Neural Network* ke dalam sistem otomatisasi menjanjikan peningkatan signifikan dalam efisiensi akuisisi informasi publikasi ilmiah. Transformer, yang dipublikasikan dalam *paper 'Attention Is All You Need'* (2017), telah direvolusi untuk mengoptimalkan pemrosesan sekuen data. Kemampuannya dalam menangani dependensi jangka panjang dalam data terbukti sangat efektif, sehingga membuatnya menjadi metode pilihan dalam berbagai aplikasi pemrosesan bahasa alami kontemporer.

LLM menunjukkan kemampuan yang mengesankan tetapi menghadapi tantangan seperti halusinasi, pengetahuan yang usang, dan proses penalaran yang tidak transparan dan tidak dapat dilacak. *Retrieval Augmented Generation (RAG)* telah muncul sebagai solusi yang menjanjikan dengan menggabungkan pengetahuan dari basis data eksternal. Pengkombinasian dengan *Retrieval Augmented Generation (RAG)* meningkatkan akurasi dan kredibilitas dari generasi respons, terutama untuk tugas-tugas yang memerlukan banyak pengetahuan, dan memungkinkan pembaruan pengetahuan yang terus-menerus serta integrasi informasi spesifik domain. *RAG* secara sinergis menggabungkan pengetahuan intrinsik *LLM* dengan repositori eksternal yang luas dan dinamis.

Sejumlah paper mengenai teknologi *chatbot* telah dikompilasi dan dianalisis dalam literatur sebelumnya. *Chatbot* telah dikembangkan oleh Khadija, Aziz, dan Nurharjadmo, yang mengintegrasikan kerangka kerja *LangChain* dan memanfaatkan kemampuan *ChatGPT* versi *OpenAI GPT-3.5* serta infrastruktur *Pinecone* untuk memfasilitasi generasi respons yang tidak hanya koheren tetapi juga kontekstual terhadap materi yang terdapat dalam dokumen *PDF* [1]. *Chatbot* telah dibangun dengan memanfaatkan teknologi *Bidirectional Encoder Representations from Transformers (BERT)*. Dengan memanfaatkan model *BERT* yang telah dilatih sebelumnya, chatbot berhasil mengekstrak informasi yang dibutuhkan pengguna dari spesifikasi konstruksi [2]. Sebuah *chatbot* pendidikan yang diberi nama *"College Enquiry Chatbot"*

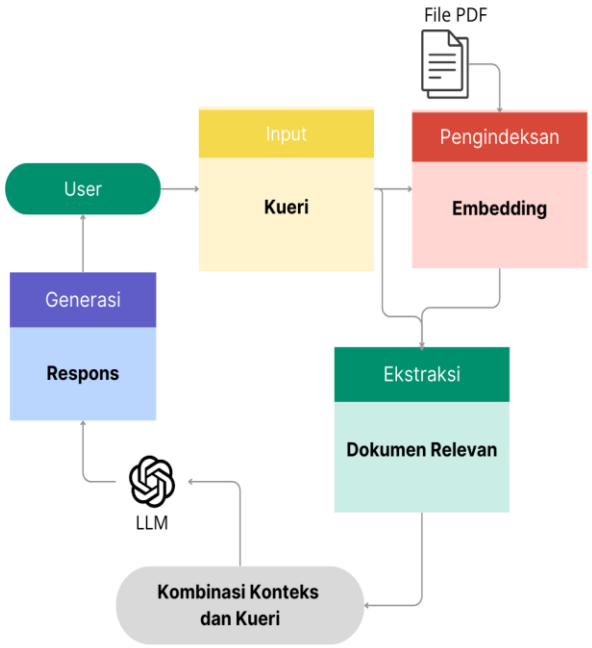


Gambar 1. Arsitektur model *transformer*.

telah dikembangkan sebagai aplikasi web yang menggunakan teknologi *Rasa*. Teknologi ini terdiri dari dua komponen utama: *Rasa Core* dan *Rasa Natural Language Understanding (NLU)*. *Rasa NLU* berfungsi untuk menginterpretasikan intensi pengguna dan mengekstrak entitas yang diperlukan dari masukan yang diberikan, sedangkan *Rasa Core* menghasilkan respons dengan membangun model probabilistik yang didukung oleh Jaringan Saraf Berulang *(RNN)*. Evaluasi dari model ini diukur melalui penerapan matriks kebingungan dan metrik kinerja seperti presisi, akurasi, dan skor F1, yang rata-rata masing-masing adalah 0,628, 0,725, dan 0,669 [3]. Kontribusi dari paper ini ialah sebagai berikut:

1. Konstruksi antarmuka *web chatbot* menggunakan *framework NextJs*.
2. Integrasi infrastruktur *Transformer OpenAI GPT-4* dan sistem *Retrieval Augmented Generation (RAG) Pinecone* ke dalam antarmuka *web*.
3. Penambahan fitur untuk mengunggah *file* berorientasi *PDF* dengan memanfaatkan sistem *Retrieval Augmented Generation (RAG)*.
4. Konstruksi dan konfigurasi *database* relasional, berkas *PDF*, dan *database* vektor.

Segmen selanjutnya dari paper ini akan mengelaborasi uraian literatur hingga ke simpulan. Diinisiasi dari segmen *chatbot* yang memberikan gambaran tentang literatur yang relevan yang menjadi teori fundamental dari paper ini. Kemudian diikuti dengan metodologi yang menguraikan pendekatan yang digunakan dalam paper. Bagian analisis hasil akan menyajikan temuan utama dari paper, dianalisis dan diinterpretasikan dalam konteks kerangka teori yang telah disinggung sebelumnya. Segmen kesimpulan menyajikan ringkasan dalam paper ini meliputi temuan utama, implikasi teoretis dan praktis, keterbatasan paper, dan saran untuk publikasi masa depan.



Gambar 2. Diagram cara kerja *RAG*.

II. CHATBOT

A. *Akuisisi Informasi*

Informasi adalah data yang telah diproses dan memberikan jawaban atas pertanyaan-pertanyaan dasar seperti siapa, apa, kapan, di mana, dan berapa banyak. Data dikumpulkan dari berbagai sumber dan diolah menjadi informasi. Informasi ini kemudian diintegrasikan ke dalam struktur pengetahuan yang ada, membantu individu atau sistem untuk memahami konteks dan relevansi data. Pengetahuan adalah informasi yang diorganisir dan dianalisis lebih lanjut untuk menjawab pertanyaan "bagaimana." Ini melibatkan pemahaman hubungan antara data dan informasi, serta kemampuan untuk menerapkannya dalam situasi praktis. Model *DIKW (Data, Information, Knowledge, Wisdom)* menggambarkan hubungan hierarkis antara data, informasi, pengetahuan, dan kebijaksanaan. Dengan demikian, akuisisi informasi adalah proses penting dalam pengembangan pengetahuan ilmiah, di mana informasi berperan sebagai jembatan antara data dan pengetahuan [4].

B. *Transformer*

Paper "Attention Is All You Need" memperkenalkan model *Transformer*, sebuah arsitektur jaringan saraf baru yang sepenuhnya mengandalkan mekanisme perhatian *(attention)*, bukan menggunakan lapisan berulang atau konvolusional. Arsitektur model *Transformer* divisualisasikan pada Gambar 1. Data masukan, biasanya dalam bentuk teks, dipecah menjadi token. Token-token ini kemudian diubah menjadi vektor melalui lapisan embedding. Untuk memasukkan informasi posisi ke dalam model yang tidak memiliki rekurensi atau konvolusi, *Transformer* menambahkan *positional encoding* ke vektor embedding. *Encoding* ini membantu model membedakan urutan token dalam urutan *input*. *Encoder* dalam *Transformer* terdiri dari N lapisan yang identik yang masing-masing berisi dua sub-lapisan: mekanisme *multi-head attention* dan *feed-forward neural*

Tabel 1.
Contoh Implementasi *Website Chatbot* Berorientasi *PDF* untuk Akuisisi Informasi dari Publikasi Akademik

Input	*Paper Automating Information Retrieval from Faculty Guidelines: Designing a PDF-Driven Chatbot powered by OpenAI ChatGPT.pdf* Apa abstraksi dari berkas tersebut?
Context	*IV. IMPLEMENTATION AND RESULT In this research, we use LangChain Framework and ChatGPT (GPT 3.5 Turbo) from Open API to get the response without fine tuning. For the vector database storage we use Pinecone [26]. In the Pinecone we set the embedding dimension in 1536 and using metric Cosine Similarity. We need API keys from Pinecone also API key from OpenAI. TABLE 1: THE EXAMPLE OF QUERY AND RESPONSE IN CHATBOT Context User wants to know what is Group Research Grant and what is the output of the grant. The* *...* *There are 2 main keywords in the question. Question 1 is "what is HGR research" and question 2 is "what are the outcomes of category A HGR research?". When looking at the original document seen in Figure 5, this chatbot can identify*

Prompt	`AI assistant is a brand new, powerful, human-like artificial intelligence.`
	`The traits of AI include expert knowledge, helpfulness, cleverness, and articulateness.`
	`AI is a well-behaved and well-mannered individual.`
	`AI is always friendly, kind, and inspiring, and he is eager to provide vivid and thoughtful responses to the user.`
	`AI has the sum of all knowledge in their brain, and is able to accurately answer nearly any question about any topic in conversation.`
	`START CONTEXT BLOCK`
	`${lastUserInput}`
	`${context}`
	`END OF CONTEXT BLOCK`
	`AI assistant will take into account any CONTEXT BLOCK that is provided in a conversation.`
	`If the context does not provide the answer to question, the AI assistant will still answer the questions and explain it with rich and vivid explanation.`

Output	Berkas tersebut dalam konteks ini menggambarkan penelitian tentang penggunaan *LangChain Framework* dan *ChatGPT (GPT 3.5 Turbo)* dari *Open API* untuk mendapatkan respons dari *chatbot* tanpa perlu penyesuaian lebih lanjut. Untuk penyimpanan *database* vektor, digunakan *Pinecone*. Penelitian ini juga menampilkan contoh bagaimana *chatbot* dapat menjawab pertanyaan yang relevan dengan konteks dokumen *PDF*. Misalnya, chatbot dapat menjelaskan tentang penelitian HGR (Hibah Grup Riset) dan hasil yang diharapkan dari kategori penelitian HGR A, berdasarkan konten yang ada dalam dokumen *PDF*. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pengguna dapat menanyakan pertanyaan terkait konteks dokumen *PDF* dan *chatbot* dapat meresponsnya dengan tepat. Kedepannya, ini dapat membantu meningkatkan interaksi antara pengguna dan *chatbot* dan membantu *chatbot* untuk membuat respon yang lebih tepat dan relevan berdasarkan konteks yang diberikan.

network. Pada sub-lapisan *attention*, model mengidentifikasi bagian mana dari kalimat yang relevan terhadap bagian lainnya, memungkinkan model untuk menangkap konteks dari keseluruhan *input* sekaligus. Untuk setiap kata dalam sebuah kalimat, lapisan *self-attention* menghitung tiga vektor *(key, value, query)*. Untuk menentukan kata-kata yang berhubungan secara kontekstual dengan sebuah kata, produk titik dari vektor *query* dipertimbangkan dengan vektor *key* dari kata-kata lain. Jaringan saraf *feed-forward* menerapkan transformasi nonlinear pada keluaran dari mekanisme *self-attention*, memperkenalkan kompleksitas dan kekuatan ekspresif pada model. *Decoder* juga terdiri dari N lapisan yang masing-masing memiliki tiga sub-lapisan: *multi-head self-attention, encoder-decoder attention*, dan *feed-forward neural network*. Lapisan ini mirip dengan *self-attention* pada *encoder* tetapi menghalangi setiap posisi untuk melihat posisi yang akan datang, yang esensial untuk pembuatan *output* secara bertahap. Lapisan ini membantu *decoder* fokus pada bagian relevan dari *input*, serupa dengan mekanisme *attention* pada *encoder*, tetapi mendapatkan *input* dari lapisan *encoder* terakhir. Setelah data melewati semua lapisan *decoder*, *output* diubah menjadi skor *logit* menggunakan lapisan linear dan diubah menjadi probabilitas menggunakan fungsi *softmax*. Ini menunjukkan probabilitas kata berikutnya dalam urutan target. *Output* dari *decoder* diinterpretasikan sebagai prediksi untuk kata berikutnya dalam teks target. Kata ini kemudian digunakan sebagai input tambahan ke *decoder* selama langkah berikutnya dari proses generasi teks. Proses ini diulangi sampai model menghasilkan token khusus yang menandakan akhir dari kalimat terjemahan atau *output* [5].

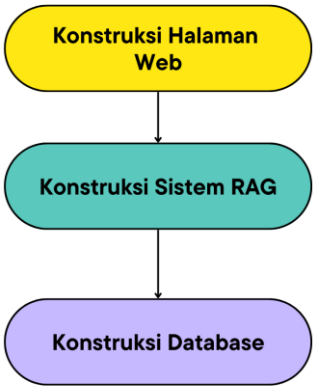
C. *Retrieval Augmented Generation (RAG)*

Retrieval-Augmented Generation (RAG) adalah sebuah teknologi yang meningkatkan model bahasa besar *(Large Language Models, LLMs)* dengan mengambil informasi relevan dari basis data eksternal melalui perhitungan kesamaan semantik. Diagram cara kerja *RAG* dapat diamati pada Gambar 2. *RAG* bekerja melalui tiga tahapan utama: pengindeksan, pengambilan, dan generasi.

1. Pengindeksan: Dokumen diproses dan dipecah menjadi bagian-bagian yang lebih kecil, kemudian diubah menjadi representasi vektor dan disimpan dalam basis data vektor.
2. Pengambilan: Ketika pengguna mengajukan pertanyaan, sistem *RAG* menggunakan model pengkodean yang sama dengan saat pengindeksan untuk mengubah pertanyaan menjadi representasi vektor, mencari kesamaan antara vektor pertanyaan dan vektor dokumen yang diindeks, dan mengambil bagian teratas yang paling relevan.
3. Generasi: Pertanyaan yang diajukan dan dokumen yang dipilih dikombinasikan menjadi prompt yang koheren yang kemudian digunakan oleh model bahasa besar untuk menghasilkan jawaban.

D. *Cosine Similarity*

Cosine similarity adalah ukuran yang digunakan untuk menentukan seberapa mirip dua vektor dalam ruang dimensi. Kesamaan kosinus antara dua vektor dihitung dengan mengambil *dot product* dari dua vektor tersebut dan membaginya dengan hasil kali panjang *(norm)* masing-masing vektor. Nilai 1 menunjukkan bahwa kedua vektor sangat mirip. Nilai 0 menunjukkan bahwa kedua vektor tidak



Gambar 3. *Flow diagram* konstruksi *website chatbot* berorientasi *PDF*.

Tabel 2.
Evaluasi Kuantitatif *Website Chatbot* Berorientasi *PDF* Menggunakan *RAGAs*

Parameter		Skor
Retrieval	*Context Recall*	0.68
(Ekstraksi)	*Context Precision*	0.79
Generation	*Answer Relevancy*	0.83
(Generasi)	*Faithfulness*	0.51
End to End Evaluation	*Answer Correctness*	0.7

memiliki kesamaan. Nilai -1 menunjukkan bahwa kedua vektor berlawanan [6]. Persamaan *cosine* dinyatakan dalam persamaan berikut.

$$Cosine\ Similarity = \frac{\overline{A}.\overline{B}}{|A||B|} \quad (1)$$

Keterangan:
$\overline{A}\ dan\ \overline{B} = dua\ vektor$
$\overline{A}.\overline{B} = dot\ product\ dari\ vektor\ \overline{A}\ dan\ \overline{B}$
$|A||B| = panjang\ (norm)\ dari\ vektor\ \overline{A}\ dan\ \overline{B}$

E. Metrik Performansi

RAGAs (Retrieval Augmented Generation Assessment) merupakan paradigma evaluatif inovatif yang diformulasikan untuk menganalisis efektivitas sistem generasi augmentatif berbasis retrieval. Paradigma evaluatif ini mengimplementasikan serangkaian indikator kuantitatif, yang mencakup namun tidak terbatas pada:

1) Faithfulness

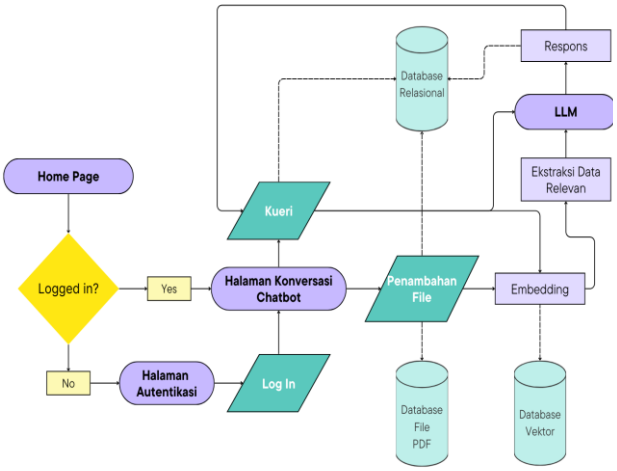
Metrik ini mengkuantifikasi konsistensi faktual antara respons yang digenerasi dan konteks yang disediakan. Kalkulasinya melibatkan analisis komparatif antara output generatif dan korpus kontekstual yang diretrieval. Hasil evaluasi dinormalisasi ke dalam rentang [0,1], dengan nilai yang lebih tinggi mengindikasikan performa superior.

$$Faithfulness = \frac{|x|}{|y|} \quad (2)$$

Keterangan:
x = Jumlah klaim dalam jawaban yang dihasilkan yang dapat disimpulkan dari konteks yang diberikan
y = Jumlah kalimat dalam *ground truth*

2) Answer Correctness

Evaluasi *answer correctness* melibatkan kuantifikasi akurasi output generatif relatif terhadap standar veritable. Proses penilaian ini mengintegrasikan analisis komparatif antara respons yang dihasilkan dan referensi otoritatif, menghasilkan metrik yang terkalibrasi dalam rentang [0,1].



Gambar 4. *User flow diagram website chatbot* berorientasi *PDF*.

Tabel 3.
Evaluasi Pre Implementasi dan Post Implementasi Integrasi Berkas *PDF* Menggunakan Parameter *Answer Correctness*

Pre Implementasi *File PDF*	Post Implementasi *File PDF*
0.5	0.7

Metodologi penilaian ini menginkorporasikan dua dimensi: semantic similarity serta factual similarity.

$$F1\ Score = \frac{|TP|}{(|TP|+0.5\times(|FP|+|FN|))} \quad (3)$$

Keterangan:
$TP = True\ Positive$
$FP = False\ Positive$
$FN = False\ Negative$

Dilakukan penghitungan rata-rata berbobot dari similaritas semantik dan similaritas faktual yang telah dihitung sebelumnya untuk menghasilkan skor akhir.

3) Answer Relevancy

Metrik evaluatif, *answer relevancy*, berfokus pada kuantifikasi koherensi antara output generatif dan stimulus input. Nilai kuantitatif yang lebih rendah mengindikasikan output yang tidak lengkap atau mengandung informasi redundan, sementara nilai yang lebih tinggi mengindikasikan kongruensi superior.

$$Answer\ Relevancy = \frac{1}{N}\sum_{i=1}^{N}\cos\left(E_{g_i}, E_o\right) \quad (4)$$

Keterangan:
N = kardinalitas himpunan pertanyaan artifisial yang diekstrapolasi, dengan nilai *default* 3 sebagai standar
E_{g_i} = vektor *embedding* dari pertanyaan artifisial i yang dihasilkan melalui proses ekstrapolasi
E_o = vektor *embedding* dari pertanyaan orisinal yang merupakan *input*

4) Context Recall

Context recall mengkuantifikasi derajat kongruensi antara konteks yang diektstraksi dengan respons teranotasi, yang diperlakukan sebagai standar veridikalitas. Metrik ini dikalkulasi berdasarkan *'ground truth'* dan *'context'*, dengan rentang nilai antara 0 dan 1, di mana nilai yang lebih tinggi mengindikasikan performa superior.

$$Context\ Recall = \frac{|x|}{|y|} \quad (5)$$

Keterangan:
x = Kalimat *ground truth* yang dapat diatribusikan ke konteks
y = Jumlah kalimat dalam *ground truth*

5) Context Precision

Context precision merupakan metrik evaluatif yang mengkuantifikasi apakah seluruh elemen relevan berdasarkan standar veridikalitas yang terdapat dalam 'konteks' memperoleh peringkat superioritas atau tidak. Dalam kondisi ideal, seluruh fragmen relevan seharusnya terposisikan pada strata teratas dalam hierarki peringkat.

$$Context\ \Pr e\ cision@K = \sum_{k=1}^{K} \frac{x}{y} \quad (6)$$

Keterangan:
K = total potongan dalam konteks

$$x = \sum_{k=1}^{K} (\Pr e\ cision@K \times v_k)$$

$v_k \in \{0,1\} = indikator\ relevansi\ pada\ peringkat\ k$
y = Jumlah total item relevan dalam K hasil teratas

$$\Pr e\ cision@K = \frac{x}{y} \quad (7)$$

Keterangan:
K = total potongan dalam konteks
$x = true\ positive@k$
$y = (true\ positive@k + false\ positive@k)$

III. METODOLOGI

Gambar 3 menggambarkan tiga tahap utama dalam konstruksi *website chatbot* yang berorientasi pada *PDF*, yaitu konstruksi halaman *web*, konstruksi sistem *Retrieval Augmented Generation (RAG)*, dan konstruksi serta konfigurasi basis data.

A. Konstruksi Halaman Web

Tahap pertama adalah konstruksi halaman web yang mencakup tiga komponen utama:

1. *Home Page*: Halaman utama yang berfungsi sebagai titik awal bagi pengguna.
2. *Authentication Page*: Halaman yang berfungsi untuk otentikasi pengguna.
3. *Chatbot Conversation Page*: Halaman di mana pengguna dapat berinteraksi dengan *chatbot*. *Chatbot* ini akan menggunakan data dari sistem *RAG* untuk menjawab pertanyaan pengguna.

B. Konstruksi Sistem Retrieval Augmented Generation (RAG)

Pinecone dikonfigurasi agar mampu menangani data dengan dimensi sebesar 1536 dan menggunakan pendekatan *cosine similarity* untuk mengukur kesamaan antarvektor. Tahap kedua adalah konstruksi sistem *RAG* yang terdiri dari beberapa langkah penting:

Memecah dokumen *PDF* menjadi bagian-bagian yang lebih kecil agar lebih mudah diolah. Program akan memotong string berdasarkan jumlah byte yang ditentukan, bukan berdasarkan jumlah karakter. String dikonversi menjadi *array byte* dalam format *UTF-8*, kemudian memotong *array byte* ini hingga batas *byte* yang spesifik. *Array byte* yang telah dipotong kemudian dikonversi kembali menjadi *string*. Program membatasi maksimal pemotongan dokumen sebesar 36000 *byte* untuk memastikan pengelolaan data yang efisien dan membatasi ukuran data yang akan diproses.

Mengubah bagian-bagian dokumen yang lebih kecil tersebut menjadi vektor. Menyimpan vektor-vektor tersebut ke dalam basis data vektor. Mengambil data yang relevan dari basis data dengan mencari kesamaan antara vektor pertanyaan dan vektor dokumen yang diindeks, dan mengambil bagian teratas yang paling relevan. Program menginisialisasi klien *Pinecone* dengan spesifikasi lingkungan dan kunci *API* yang diambil dari variabel lingkungan. Program akan berfokus pada dokumen dengan *matching score* di atas 0.7, yang menandakan tingkat relevansi yang tinggi. Setiap dokumen yang memenuhi syarat ini kemudian diproses untuk mengumpulkan teksnya (dan nomor halaman, jika relevan, meskipun tidak langsung ditampilkan dalam fungsi). Teks-teks ini digabungkan menjadi satu *string*, dipisahkan oleh baris baru, dan dibatasi panjangnya hingga 3000 karakter untuk memastikan respons yang *compact* dan terfokus.

Menggabungkan prompt dari pengguna dengan data relevan dari dokumen untuk diolah oleh *Large Language Model (LLM)*, yang kemudian memberikan respons kepada pengguna. *OpenAI API* akan menangani prompt yang telah dikombinasi dan merespons menggunakan model *GPT-4* dari *OpenAI*.

C. Konstruksi dan Konfigurasi Basis Data

Tahap terakhir adalah konstruksi dan konfigurasi basis data yang mencakup beberapa jenis basis data: *Relational Database*: Basis data relasional untuk menyimpan informasi terstruktur seperti metadata dokumen dan informasi pengguna. Kombinasi *Neon* sebagai *database backend* dan *DrizzleORM* sebagai lapisan interaksi data menawarkan struktur yang efisien dan *user friendly* untuk pengembangan aplikasi, khususnya untuk *website chatbot* yang dibangun.

PDF File Database: Basis data yang digunakan untuk menyimpan file *PDF* mentah. *Amazon S3 (Simple Storage Service)* adalah layanan penyimpanan objek yang disediakan oleh *Amazon Web Services (AWS)* dan sering digunakan untuk menyimpan dan mengambil data dalam jumlah besar. Agar dapat menggunakan *platform* ini, maka diperlukan inisialisasi *bucket*, konfigurasi *bucket policy* dan *CORS* terlebih dahulu. *Bucket Policy* adalah salah satu alat yang digunakan dalam *AWS S3* untuk mengelola izin akses ke *bucket* dan objek yang ada di dalamnya. *Bucket policy* memungkinkan akses baca dan hapus secara terbuka kepada semua objek di *bucket final-project-001*. *CORS (Cross-Origin Resource Sharing)* adalah mekanisme yang memungkinkan banyak sumber daya (misalnya, *font*, *JavaScript*, dll.) pada halaman *web* untuk diminta dari domain yang berbeda dari domain dari mana sumber daya pertama kali di-*serve*. Untuk mengakses *AWS S3* dalam pengembangan *website*, dibutuhkan *API key*, yang terdiri dari *AWS Access Key ID* dan *AWS Secret Access Key*.

Vector Database: Basis data untuk menyimpan representasi vektor dari dokumen *PDF* yang telah diolah. *Pinecone* adalah *platform database* yang dirancang khusus untuk memudahkan pengelolaan dan penerapan sistem berbasis vektor pada skala besar. Untuk mengakses *pinecone* dalam konstruksi *website*, variabel *environment* dan *API key* dibutuhkan untuk keamanan dan fleksibilitas konfigurasi.

D. User Flow Diagram Website Chatbot Berorientasi PDF

Pengguna memulai di laman utama *website*. Jika pengguna sudah *log in*, pengguna akan diarahkan ke halaman percakapan *chatbot*. Jika pengguna belum *log in*, pengguna akan diarahkan ke halaman untuk *log in* terlebih dahulu. Pada halaman percakapan *website*, pengguna dapat berinteraksi dengan *chatbot*. Pengguna perlu menambahkan *file PDF* terlebih dahulu yang kemudian diunggah ke *Database File PDF*. *File PDF* yang diunggah diproses dan dibagi menjadi bagian-bagian yang lebih kecil atau *"chunks"* yang relevan untuk *query* yang dibuat oleh pengguna. Bagian-bagian dokumen yang telah diproses kemudian diubah menjadi vektor melalui proses *embedding* dan disimpan dalam *Database* Vektor. Pengguna kemudian dapat mengajukan pertanyaan atau permintaan. Saat pengguna membuat pertanyaan, chatbot mengambil pertanyaan tersebut dan menggabungkannya dengan data yang relevan dari *PDF* untuk membuat prompt yang akan diproses oleh model bahasa *(LLM)*. Model bahasa menggunakan data tersebut untuk mencari dan mengambil informasi yang relevan sebagai respons kepada pengguna. Setelah data yang relevan ditemukan dan diolah, chatbot kemudian menghasilkan respons yang dikirimkan kembali ke pengguna. User flow diagram untuk website chatbot yang difokuskan pada pengelolaan dokumen PDF dapat diamati pada Gambar 4.

IV. ANALISIS HASIL

Chatbot telah berhasil melakukan ekstraksi esensi dari dokumen akademis dan mampu memberikan respons terhadap pertanyaan yang berkaitan dengan dokumen *PDF*. Diuraikan suatu contoh penerapan *chatbot* untuk akuisisi informasi konten publikasi akademik melalui Tabel 1.

A. Deskripsi Dataset dan Persiapan

Untuk kebutuhan evaluasi kuantitatif, suatu korpus data telah dikonstruksi, terdiri dari 100 entitas data berbentuk pilihan ganda. Setiap entitas data menginkorporasikan empat komponen esensial: *question*, *ground truth*, *context*, dan *answer* dari sistem interaksi berbasis kecerdasan artifisial. Korpus data ini telah mengalami spesialisasi dalam domain epistemologi filosofis, dengan sumber perolehan data berasal dari repositori digital *global.oup.com* yang berafiliasi dengan institusi akademik *Oxford University Press*. Tujuan fundamental dari konstruksi *dataset* ini adalah untuk memfasilitasi evaluasi kuantitatif terhadap performa sistem *Retrieval-Augmented Generation (RAG)*. Metodologi evaluatif yang diimplementasikan mengadopsi *framework* dan metrik-metrik yang terspesifikasi dalam paradigma *RAGAs (Retrieval-Augmented Generation Assessments)*.

Sebagai bagian dari evaluasi kualitatif intersubjektif, disusun suatu survei pengguna guna mengevaluasi kinerja *chatbot* dalam memproses dan merespons pertanyaan yang berkaitan dengan konten publikasi akademik yang berfokus pada *PDF* akademik, serta untuk menilai tingkat kemudahan penggunaan dan kualitas antarmuka pengguna. Selain itu, survei ini bertujuan untuk mengumpulkan *feedback* dari pengguna guna mendapatkan saran untuk meningkatkan kinerja chatbot di masa mendatang. Kegiatan survei melibatkan lima responden yang merupakan mahasiswa. Survei terdiri dari 11 pertanyaan, dengan 9 di antaranya berbentuk pilihan ganda dan 2 lainnya berupa pertanyaan perbaikan dan saran. Studi ini melibatkan sejumlah lima mahasiswa sebagai partisipan dalam proses evaluasi intersubjektif, yang bertujuan untuk menganalisis dan menginterpretasikan data melalui perspektif kolektif. Meskipun ukuran sampel relatif kecil, pendekatan ini memungkinkan analisis yang lebih mendalam per partisipan, yang dapat sangat berharga dalam konteks iterasi desain *chatbot* yang cepat dan efisien.

B. Evaluasi Kuantitatif

Evaluasi ini mengacu pada proses penilaian kinerja sistem *Retrieval Augmented Generation (RAG)* dengan menggunakan metrik dan metode menggunakan *RAGAs (Retrieval Augmented Generation Assessment)* yang dapat diukur secara numerik. *Framework RAGAs* menyediakan serangkaian alat dan metrik untuk mengukur berbagai aspek kinerja *RAG* secara spesifik, seperti akurasi, relevansi, dan efisiensi dari jawaban yang dihasilkan [7]. Metrik evaluasi yang sering diadopsi seperti *ROUGE* dan *BLEU* tidak memiliki relevansi dalam konteks ini. Hal ini dikarenakan *ROUGE* secara intrinsik berorientasi pada penilaian tugas peringkasan, sementara *BLEU* didedikasikan untuk evaluasi tugas penerjemahan bahasa. Evaluasi komprehensif terhadap sistem interaksi berbasis kecerdasan artifisial telah menghasilkan profil performa multidimensional yang dapat diamati pada Tabel 2.

C. Evaluasi Pre Implementasi dan Post Implementasi Integrasi Berkasi PDF

Pengujian dilakukan dengan menggunakan parameter *answer correctness* dari *RAGAs (RAG Assessment)*. Parameter ini digunakan untuk mengukur seberapa akurat jawaban yang diberikan oleh sistem sebelum dan sesudah integrasi berkas *PDF*. Pada tahap pre-implementasi, tingkat keakuratan jawaban berada pada angka 0.5, yang berarti hanya setengah dari jawaban yang diberikan adalah benar. Setelah dilakukan integrasi berkas *PDF*, pada tahap post-implementasi, tingkat keakuratan jawaban meningkat menjadi 0.7, menunjukkan adanya peningkatan akurasi sistem dalam memberikan jawaban yang benar sebesar 20%. Evaluasi ini menunjukkan bahwa integrasi berkas *PDF* berdampak positif terhadap performa sistem dalam konteks keakuratan jawaban yang dihasilkan. Hasil evaluasi pre implementasi dan post implementasi integrasi berkas *PDF* dapat diamati pada Tabel 3.

V. KESIMPULAN/RINGKASAN

Adapun kesimpulan yang dapat diekstraksi dari paper ini adalah sebagai berikut: (1) Integrasi model *Transformer* seperti *GPT-4* dengan sistem *chatbot* mampu secara efektif memahami dan mengekstrak informasi dari dokumen *PDF*. (2) Sistem *Retrieval Augmented Generation (RAG)* dengan memanfaatkan *Pinecone* yang digunakan dalam *chatbot* ini meningkatkan akurasi dalam mengekstrak informasi spesifik dari dokumen *PDF*. (3) Konstruksi *website chatbot* berorientasi *PDF* meningkatkan pengalaman pengguna dengan menyediakan akses cepat dan efisien ke informasi yang diperlukan. (4) Penggunaan *chatbot* berorientasi *PDF* secara signifikan mengurangi waktu yang diperlukan untuk mengekstrak dan menganalisis informasi dari dokumen *PDF*

daripada proses akuisisi informasi yang dilakukan secara manual.

Beberapa saran diberikan yang diharapkan dapat menjadi landasan bagi pengembangan dan peningkatan paper ini, sehingga dapat lebih relevan dan bermanfaat dalam konteks yang lebih luas yaitu: (1) Peningkatan dalam kemampuan *chatbot* untuk mengelola berbagai jenis *input*, termasuk format *file* yang berbeda seperti *Word*, *Excel*, dan sebagainya. (2) Penambahan fitur *input* suara dan pengenalan gambar sehingga pengaksesan informasi dapat dilakukan dari berbagai sumber dan format, meningkatkan kualitas layanan yang diberikan kepada pengguna.

DAFTAR PUSTAKA

[1] M. A. Khadija, A. Aziz, and W. Nurharjadmo, "Automating information retrieval from faculty guidelines: designing a PDF-Driven Chatbot powered by OpenAI ChatGPT," in *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Bandung, 2023, pp. 394--399.

[2] J. Kim, S. Chung, S. Moon, and S. Chi, "Feasibility study of a BERT-based question answering chatbot for information retrieval from construction specifications," in *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Kuala Lumpur, 2022, pp. 0970--0974.

[3] S. Meshram, N. Naik, V. Megha, T. More, and S. Kharche, "College enquiry chatbot using rasa framework," in *Asian Conference on Innovation in Technology (ASIANCON)*, Pune, 2021, pp. 1--8.

[4] B. Bosancic, "Information in the knowledge acquisition process," *J. Doc.*, vol. 72, no. 5, pp. 930--960, 2016.

[5] A. Vaswani, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, California, 2017.

[6] D. Gunawan, C. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," *J. Phys. Conf. Ser.*, vol. 978, 2018.

[7] M. Dean, R. R. Bond, M. F. McTear, and M. D. Mulvenna, "ChatPapers: An AI chatbot for interacting with academic research," in *Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, Letterkenny, 2023, pp. 1--7.