Penerapan QnA Chatbot dengan Metode Mixtral-8x7B dan RAG: Studi Kasus pada Dokumen Kesehatan Tentang Parenting

Nurul Najwa Sabilla, Sidik Riffani
Teknik Informatika, Universitas Pelitas Bangsa
Jalan Inspeksi Kalimalang, Tegal Danas, No. 9, Kabupaten Bekasi, Indonesia
{nurulnazwasabilla, sidikriffani481}@gmail.com

Abstract

Masa anak usia dini merupakan tahap krusial dalam tumbuh kembang anak, yang sangat dipengaruhi oleh kualitas pola asuh orang tua. Namun, masih banyak orang tua di Indonesia yang minim pengetahuan tentang pengasuhan, terutama karena faktor usia muda dan kurangnya akses informasi. Hal ini berdampak pada masalah gizi, pertumbuhan, hingga kesehatan mental anak. Penelitian ini bertujuan untuk mengevaluasi efektivitas model bahasa *Mixtral-8x7B-Instruct-v0.1* dalam memahami dan mengolah dokumen *parenting* menggunakan pendekatan *Retrieval-Augmented Generation* (RAG). Data *parenting* diperoleh melalui *web scraping* dari situs Hello Sehat, kemudian diproses dengan metode *chunking*, *embedding* menggunakan model *all-mpnet-base-v2*, dan disimpan dalam *Chroma DB* sebagai vektor. Sistem dibangun dengan menggabungkan *retriever*, *prompt template*, dan language model untuk menjawab pertanyaan berbasis konteks. Hasil pengujian terhadap pertanyaan seperti "Kapan Anak Harus Mendapatkan Vaksin Hepatitis?" menunjukkan bahwa sistem dapat memberikan jawaban yang ringkas dan relevan. Evaluasi menggunakan *BERTScore* menunjukkan nilai *F1-Score* sebesar 0.8242%, menandakan kesamaan semantik tinggi dengan referensi. Temuan ini membuktikan bahwa integrasi *Mixtral* dan RAG efektif dalam memproses dokumen parenting serta berpotensi mendukung edukasi pengasuhan di Indonesia.

1 Pendahuluan

Masa balita merupakan fase krusial dalam kehidupan anak karena pada tahap ini terjadi percepatan yang signifikan dalam proses pertumbuhan dan perkembangan. Berdasarkan data dari Organisasi Kesehatan Dunia (WHO), sekitar 7,3% balita mengalami kekurangan gizi, 5,9% mengalami kelebihan berat badan, dan 21,9% mengalami hambatan pertumbuhan [1]. Di negara-negara berkembang, termasuk Indonesia, laju pertumbuhan dan perkembangan anak cenderung lebih lambat dibandingkan dengan negara maju, yang sebagian besar disebabkan oleh asupan gizi yang tidak mencukupi. Dengan penanganan yang tepat sejak dini, anak-anak Indonesia memiliki peluang untuk tumbuh dan berkembang secara optimal [2].

Perkembangan anak sangat dipengaruhi oleh keterlibatan orang tua. Kualitas pola pengasuhan memiliki peran penting dalam mendukung pertumbuhan dan perkembangan anak. Pola asuh yang positif, disertai dengan komunikasi yang efektif, dapat mendorong perkembangan anak secara optimal. Sebaliknya, pengasuhan yang tidak memadai dapat memberikan dampak negatif terhadap proses

tumbuh kembang anak. Upaya intervensi dalam pengasuhan dapat membantu meningkatkan kualitas interaksi orang tua dengan anak, memperbaiki kondisi kesehatan anak, serta memperkuat hubungan emosional yang sehat antara orang tua dan anak. Salah satu hambatan utama yang dihadapi orang tua adalah minimnya pengetahuan tentang pengasuhan dan kurangnya persiapan menjadi orang tua. Hal ini sering terjadi karena sebagian orang tua menikah di usia muda dan belum memiliki pengetahuan serta keterampilan pengasuhan yang memadai [2].

Salah satu fase *parenting* dalam siklus pendidikan juga terletak pada masa anak usia dini, yaitu pada rentang usia 0 hingga 8 tahun, sebagaimana ditegaskan dalam pendekatan *Developmentally Appropriate Practices* (DAP). Pada tahap ini, anak mengalami pertumbuhan dan perkembangan fisik serta mental yang sangat pesat, sehingga memerlukan pola asuh dan intervensi pendidikan yang sesuai dengan tahapan perkembangannya [3]. Seiring berjalannya waktu dan perkembangan zaman, muncul kasus-kasus yang terjadi antara orang tua dan anak. Temuan ini diperoleh dari Riset Kesehatan Dasar (Riskesdas) tahun 2018, menunjukkan bahwa remaja berusia di atas 15 tahun mengalami peningkatan gangguan mental sebesar 9,8%[4].

Permasalahan ini semakin kompleks ketika situasi keluarga berada dalam kondisi khusus seperti perceraian atau status orang tua tunggal *(single parent)*, yang sering kali berdampak pada kestabilan emosional anak. Studi Hetherington dan Kelly mengungkapkan bahwa sekitar 25% anak hasil perceraian mengalami masalah sosial dan emosional serius di masa dewasa awal, dibandingkan dengan 10% pada anak dari keluarga utuh. Amato (2005) juga menyebutkan bahwa meskipun anak dalam keluarga orang tua tunggal dapat berfungsi dengan baik, mereka cenderung mengalami hambatan dalam perkembangan sosial dan akademik [5].

Maraknya permasalahan dalam hubungan orang tua dan anak menunjukkan pentingnya edukasi *parenting* untuk mencegah konflik, gangguan mental, dan kekerasan terhadap anak. Proses pengasuhan membutuhkan kedisiplinan, baik dari orang tua maupun anak. Istilah “disiplin” berasal dari bahasa Latin *“disciplina”* yang berarti mendidik secara positif dan konstruktif. Konsep dasar *positive discipline* dalam *parenting* dikembangkan untuk menghindari praktik hukuman atau kekerasan dalam mendidik anak [4]. Pendekatan *positive discipline*, sebagaimana dijelaskan oleh Nasri, Adnan, dan Sulvinajayanti, mencakup beberapa aspek penting yang bertujuan untuk menciptakan pola asuh yang lebih konstruktif dan empatik. Aspek-aspek tersebut meliputi peningkatan pengetahuan orang dewasa tentang perkembangan anak. Seluruh aspek tersebut saling berkaitan dalam menciptakan lingkungan pengasuhan yang positif dan mendukung tumbuh kembang anak secara optimal [6].

Seiring dengan perkembangan teknologi, penyampaian informasi untuk tujuan edukasi kepada masyarakat semakin beragam. Salah satu teknologi yang kini banyak diterapkan sebagai media pencarian informasi adalah *chatbot* [7]. *Mixtral-8x7B* diperkenalkan sebagai model bahasa yang dirancang dengan efisiensi dan kinerja tinggi untuk memenuhi tantangan dalam *Natural Language Processing* (NLP) [8]. Mixtral didasarkan pada arsitektur *transformer*, Hal yang membedakan *mixtral* dapat memproses konteks sepanjang 32.000 token secara penuh [9]. *Mixtral* terintegrasi dengan metode *Retrieval-Augmented Generation* (RAG), menggabungkan pendekatan pengambilan dan generasi untuk memaksimalkan efektivitas. Pendekatan ini memungkinkan sistem untuk membandingkan *output* yang dihasilkan dengan satu atau lebih referensi sebelum menghasilkan jawaban yang akurat.

Penelitian ini bertujuan untuk mengevaluasi efektivitas model *Mixtral-8x7B* dalam mengolah data *parenting*. *Parenting* dipilih sebagai kumpulan data utama karena relevansinya dengan pola asuh di Indonesia. Dengan menggabungkan model *mixtral* dengan metode RAG, penelitian ini diharapkan dapat berkontribusi secara signifikan terhadap pengembangan teknologi AI dalam melacak, menganalisis, dan memahami dokumen *parenting* secara lebih efisien.

2 Kajian Teori

2.1 Artificial Intelligence (AI)

Dalam kecerdasan buatan *(artificial intelligence)*, istilah *problem solving* dan *search* digunakan untuk menggambarkan berbagai ide yang berkaitan dengan cara mengambil kesimpulan, membuat rencana, menggunakan logika, serta membuktikan suatu pernyataan. Ide-ide ini banyak digunakan dalam

program seperti pengenalan bahasa alami, pencarian informasi, pemrograman otomatis, robot, analisis teks, permainan komputer, sistem pakar, dan pembuktian matematika. Kecerdasan buatan juga dipelajari sebagai bagian dari agen cerdas, yaitu sistem yang bisa menerima informasi dari lingkungan sekitarnya dan merespons dengan tindakan tertentu [10].

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan salah satu cabang ilmu *Artificial Intelligence* (AI) yang berfokus pada pengolahan bahasa alami. Bahasa alami sendiri adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Bahasa yang diterima oleh komputer harus diproses dan dipahami terlebih dahulu agar maksud dari pengguna bisa dipahami dengan baik oleh kompute

Natural Language Processing (NLP) adalah sebuah bagian dari penelitian dan aplikasi yang mengkaji bagaimana komputer dapat digunakan untuk mengerti dan memanipulasi bahasa alami yang berupa teks atau ucapan untuk hal-hal yang berguna. Manipulasi teks telah dikenal sebagai sebuah bidang penelitian yang penting dalam NLP. Sebuah sistem NLP yang mengolah teks dimulai dengan analisis morfologi. Teks dikonversi, dalam kueri atau dokumen, untuk mendapatkan varian morfologi kata-kata yang terlibat. Pengolahan leksikal dan sintaktis melibatkan pemanfaatan kamus untuk menentukan karakteristik dari kata- kata, pengenalan part-of-speech, menentukan kata-kata dan frasa, serta untuk penguraian kalimat [11].

2.3 Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) merupakan pendekatan baru yang menggabungkan *pretrained language model* dengan mekanisme retrieval untuk meningkatkan kinerja pada tugas-tugas yang membutuhkan pengetahuan kompleks [12]. Dua unsur utama dari RAG adalah sistem pencarian *(retrieval system)* dan *neural language model*. Sistem pencarian mengekstrak informasi yang relevan dari korpus atau basis data yang besar dengan menggunakan teknik-teknik seperti pencocokan *keyword* dan pencarian semantik untuk menemukan data eksternal yang relevan secara kontekstual. Dalam RAG, informasi yang diambil diintegrasikan dengan pengetahuan *neural language model* yang telah dilatih sebelumnya, sehingga menciptakan konteks yang lebih komprehensif. Integrasi ini meningkatkan relevansi dan keakuratan konten yang dihasilkan dengan memasukkan informasi spesifik dan terkini dari sumber eksternal [13].

Fine-tuning merupakan proses pelatihan *model pretrained* lebih lanjut pada *dataset* spesifik untuk menyesuaikannya dengan tugas atau domain tertentu. Namun, untuk melakukan *fine-tuning* LLM seringkali dibutuhkan daya komputasi yang besar untuk pelatihan dan pengoperasian modelnya. Meskipun bukan metode *fine-tuning* konvensional, untuk tugas NLP seperti *question-answering* atau *text-generation*, metode RAG dapat menjadi alternatif untuk memberikan pengetahuan baru kepada LLM [13].

2.4 Transformer-Based

Model *transformer* banyak digunakan dalam NLP karena mampu memahami hubungan antar kata dengan cepat dan efisien. Salah satu model terbaru adalah *Mixtral-8x7B* dari Mistral AI, yang menggunakan arsitektur *Sparse Mixture-of-Experts* (SMoE). Model ini hanya mengaktifkan sebagian kecil parameternya untuk setiap token, sehingga lebih hemat komputasi namun tetap akurat.

Mixtral adalah model *decoder-only* yang dilatih dengan data multibahasa dan mampu menangani konteks panjang. Kinerjanya terbukti unggul, bahkan melampaui Llama 2 70B dan GPT-3.5 di beberapa benchmark, menjadikannya cocok untuk tugas-tugas NLP modern seperti retrieval-augmented generation (RAG) [9].

2.5 Hugging Face Ecosystem

Hugging Face ecosystem merujuk pada kumpulan alat*, library*, model, dan referensi terkait NLP yang dikembangkan oleh *Hugging Face* [14].

2.6 LangChain

LangChain adalah sebuah *framework* yang dirancang untuk meningkatkan kemampuan *Large Language Model* (LLM) dengan menangani kekurangannya. LLM, meskipun canggih, memiliki kendala seperti panjang *context window* yang terbatas dan ketidakmampuan untuk berinteraksi dengan sumber data eksternal. LangChain mengurangi masalah ini melalui berbagai modul seperti *Model I/O*, *Retrieval*, *Chains*, *Memory*, *Agent*, dan *Callback*. Modul-modul ini secara kolektif memperluas fungsionalitas LLM, membuatnya lebih fleksibel dan dapat beradaptasi dengan berbagai tugas dan sumber data yang lebih luas [15].

2.7 BERTScore

BERTScore adalah metode untuk mengevaluasi kualitas teks dengan bantuan NLP. Metode ini memakai model *BERT* untuk membandingkan teks hasil dengan teks acuan. Caranya dengan menghitung kesamaan kosinus antara kata-kata dalam kedua teks. Nilai *BERTScore* berkisar antara 0 hingga 1, di mana 1 berarti sangat mirip, dan 0 berarti tidak mirip sama sekali. Hasil penilaian ini mencakup tiga metrik utama: *precision*, *recall*, dan *F1-score*, yang semuanya dihitung berdasarkan kemiripan konteks antar kata [8].

a. *Precision*

Precision diperoleh dengan cara mencocokan *(pairwise cosine similarity)* token-token dalam jawaban sistem (kandidat) dengan token-token referensi. Berikut adalah rumus perhitungan *precision*, pada persamaan (1) [16].

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}j\in\hat{x}} max\ x_i\ ^T\hat{x}j \quad (1)$$

b. *Recall*

Recall diperoleh dengan cara mencocokan *(pairwise cosine similarity)* token-token dalam kalimat referensi dengan token-token kalimat kandidat. Berikut persamaan (2) untuk rumus *recall* [16].

$$R_{BERT} = \frac{1}{|\hat{x}|}\sum_{xj\in\hat{x}} max\ x_i\ ^T\hat{x}j \quad (2)$$

c. *F1-Score*

F1-score merupakan kombinasi *precision* dan *recall*. Berikut rumus perhitungan *f1-score* ditunjukkan pada persamaan (3) [16].

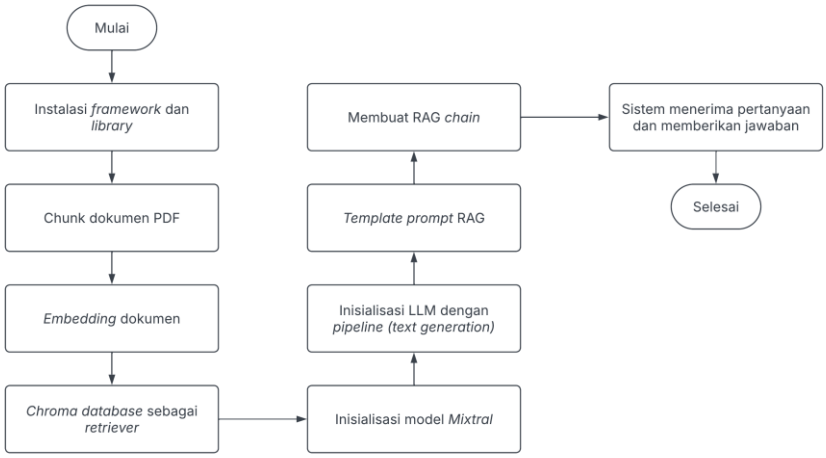
$$F_{BERT} = 2\ \frac{P_{BERT}\cdot R_{BERT}}{P_{BERT}\cdot R_{BERT}} \quad (3)$$

3 Solusi Usulan

Penelitian ini dimulai dengan menyiapkan dan mengkonfigurasi perangkat lunak yang diperlukan dengan tahapan yang ditunjukkan pada Gambar 1. Tahapannya termasuk menginstal keranga kerja *LangChain* dan beberapa pustaka pendukung, seperti *sentence-transformers* dan *huggingface_hub* untuk integrasi model *Mixtral-8x7B* melalui *LangChain*, dan GPU untuk mempercepat pelatihan dan inferensi. LangChain adalah kerangka kerja utama yang mengintegrasikan model bahasa besar *(Mixtral-8x7B)* dengan komponen lain, seperti *retriever* dalam sistem RAG. Setelah instalasi, model dan *tokenizer Mixtral-8x7B* diinisialisasi untuk memproses teks. *Tokenizer* memecah input teks menjadi token yang dapat diproses oleh *Mixtral-8x7B* sebagai inti dari sistem yang bertanggung jawab untuk memahami pertanyaan, menganalisis konteks, dan menghasilkan jawaban. Himpunan data kemudian diproses menggunakan metode *Retrieval-Augmented Generation* (RAG), di mana teks dipecah menjadi potongan-potongan dan diubah menjadi representasi numerik *(embedding)* untuk memfasilitasi pencarian semantik oleh *retriever*.

Embedding yang dihasilkan disimpan dalam *database Chroma*, yang berfungsi sebagai *retriever* untuk mencocokkan data yang relevan berdasarkan kueri pengguna. Sebuah alur teks dibangun untuk mengintegrasikan konteks yang diambil oleh *database* dengan kemampuan generatif *Mixtral-8x7B*. *LangChain* juga mengatur alur kerja sistem menggunakan *template prompt*, memastikan jawaban yang

dihasilkan diformat dan dikontekstualisasikan agar sesuai dengan kebutuhan pengguna. Fungsi utama *Mixtral-8x7B* adalah untuk membuat jawaban berbasis konteks yang berasal dari *retriever*. Pada saat yang sama, RAG memastikan relevansi dengan mengambil potongan teks yang sesuai dari himpunan data. Ini melibatkan penggabungan dua kemampuan utama: mengambil data yang relevan dan menghasilkan jawaban untuk respons akhir.

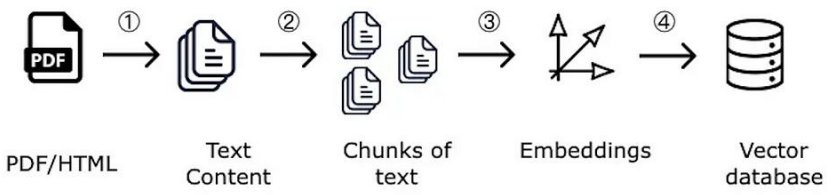


Gambar 1. Alur Penelitian

4 Hasil Eksperimen dan Pengujian

4.1 Persiapan Data

Dataset dalam penelitian ini diperoleh dari situs resmi Hello Sehat pada bagian *parenting* (https://hellosehat.com/parenting), yang memuat berbagai artikel kesehatan anak yang ditulis oleh tenaga medis profesional, termasuk dokter dan ahli kesehatan. Proses pengumpulan data dilakukan melalui teknik *web scraping*, yaitu dengan mengekstraksi konten dari halaman-halaman artikel dan menyimpannya dalam format PDF. Dataset ini mencakup kumpulan artikel *parenting* yang dikelompokkan ke dalam lima kategori utama, yaitu: (1) bayi, (2) anak usia 1–5 tahun, (3) anak usia 6–9 tahun, (4) remaja, dan (5) kesehatan anak. Data dalam format PDF memiliki *record format* kategori, judul, dan isi artikel.



Gambar 2. *Preprocessing chunk* dalam RAG

Setelah, teks diperoleh, maka akan dilakuan proses pemecahan dokumen menjadi bagian-bagian kecil *(chunkin)* menggunakan metode *RecursiveCharacterTextSplitter*. Teknik ini membagi dokumen menjadi potongan teks dengan ukuran tertentu, dalam hal ini 1.000 karakter dengan tumpang tindih (overlap) sebesar 200 karakter

4.2 *Embedding*

Gambar 2 diatas melakukan proses *embedding* data menggunaan *LLM* dengan model dari *Huggungface* yaitu *sentence-transformers/all-mpnet-base-v2*. Model all-mpnet-base-v2 adalah model yang digunakan untuk mengubah kalimat atau paragraf menjadi bentuk angka-angka (disebut vektor) yang berukuran 768. Model ini dirancang untuk digunakan sebagai *encoder* kalimat dan paragraf pendek. Dengan teks masukan, model ini menghasilkan vektor yang menangkap informasi semantik. Vektor kalimat ini dapat digunakan untuk tugas pengambilan informasi, pengelompokan, atau kesamaan kalimat.

4.3 Konfigurasi Vectore Database dan Retriever

Peneliti menggunakan *Chroma DB* sebagai *vector store* untuk melakukan pencarian dokumen yang relevan berdasarkan pertanyaan. Dengan fungsi *similarity_search_with_score()*, sistem tidak hanya mengembalikan potongan teks yang sesuai, tetapi juga memberikan skor kemiripan berdasarkan *cosine similarity*. Skor ini menggambarkan seberapa dekat makna antara pertanyaan dan isi dokumen.

Sebagai eksperiman, pertanyaan **"Kapan Anak Harus Mendapatkan Vaksin Hepatitis?"** menghasilkan beberapa potongan teks dengan skor kemiripan bervariasi. Salah satu cuplikan yang relevan menyatakan bahwa vaksin hepatitis A dan B merupakan perlindungan paling efektif bagi anak dan harus diberikan sesuai jadwal. Skor kemiripan pada hasil ini mencapai 0.344, menunjukkan relevansi moderat terhadap pertanyaan. Tabel 1 berikut merangkum beberapa hasil pencarian beserta nilai skor kemiripannya.

Tabel 1: Hasil Similarity Search: *"Kapan Anak Harus Mendapatkan Vaksin Hepatitis?"*

Ringkasan Dokumen	Skor
Vaksin hepatitis efektif untuk melindungi anak, perlu diberikan A dan B secara lengkap...	0.344
Dosis awal vaksin hepatitis B diberikan dalam 24 jam setelah lahir...	0.353
Remaja yang belum divaksin saat bayi perlu divaksin ulang...	0.356
Edukasi dan sanitasi mendukung pencegahan hepatitis di sekolah...	0.361
CDC anjurkan vaksin hepatitis B untuk semua bayi baru lahir...	0.378
Jadwal vaksin hepatitis B: 4 kali suntikan dan 1 booster...	0.391
Vaksin hepatitis A dan B kadang digabung, perlindungan bisa seumur hidup...	0.398
Dosis awal vaksin sangat penting dalam 24 jam pascalahir...	0.405
Efek samping ringan dapat terjadi setelah vaksin hepatitis B...	0.407

Setelah dokumen dimasukkan ke dalam *vector database (Chroma DB)*, peneliti mengonfigurasi *retriever* menggunakan *as_retriever()* untuk memudahkan pencarian dokumen yang relevan. Dengan pendekatan *similarity search*, sistem dapat menemukan potongan teks paling sesuai dengan pertanyaan. untuk pertanyaan *"Kapan Anak Harus Mendapatkan Vaksin Hepatitis?"*, sistem menghasilkan jawaban seperti Tabel 2.

Tabel 2: Hasil Jawaban

Hasil Jawaban Pertama
Vaksinasi hepatitis adalah bentuk perlindungan paling efektif untuk mencegah infeksi virus hepatitis. Anak perlu mendapat vaksin hepatitis A dan B untuk melindunginya dari infeksi virus tersebut. Jadwal vaksinasi hepatitis anak perlu diberikan secara lengkap agar perlindungannya optimal.
Hasil Jawaban Kedua
Sebaiknya anak menerima 3 dosis vaksin hepatitis B saja. Bila saat bayi belum menerima vaksin hepatitis B, ketika remaja ia wajib mendapatkan vaksin. Jadwal vaksin hepatitis untuk anak ini sebanyak 3 sampai 4 kali dengan masing-masing dosis berkisar antara 5–20 mg. Perlindungan vaksin hepatitis ini dapat bertahan hingga jangka panjang.

Hasil ini menunjukkan bahwa *retriever* membantu model menemukan informasi yang relevan secara semantik sebelum dijawab oleh *language model*.

4.4 Language Model

Inisialisasi LLM dengan digunakan dalam penelitian ini adalah *Mixtral-8x7B-Instruct-v0.1*, sebuah model bahasa instruksi berbasis *Mixture of Experts* dari Mistral AI. Dengan hanya dua dari delapan expert yang aktif setiap kali inferensi, model ini efisien namun tetap akurat. Mixtral telah di-*fine-tune* untuk memahami instruksi, sehingga cocok untuk tugas tanya jawab dalam sistem RAG.

4.5 Template untuk Prompt RAG

Pada tahap ini, peneliti menyusun *prompt template* sebagai bagian dari skema *Retrieval-Augmented Generation (RAG)* untuk membimbing model bahasa dalam menghasilkan jawaban yang relevan dan sesuai konteks. Template ini dirancang agar model hanya menjawab berdasarkan konteks dokumen yang diberikan oleh *retriever*. Jika jawaban tidak ditemukan dalam konteks tersebut, model diarahkan untuk menyatakan tidak tahu, guna menghindari halusinasi. Struktur prompt juga membatasi jawaban maksimal tiga kalimat agar tetap ringkas, serta diakhiri dengan frasa *"thanks for asking!"* untuk menjaga konsistensi *output*. Template ini dibangun menggunakan modul *PromptTemplate* dari *LangChain*.

4.6 Membuat RAG Chain (Retriever + Prompt + LLM)

Untuk menghasilkan jawaban berdasarkan dokumen yang relevan, digunakan *pipeline* RAG *(Retrieval-Augmented Generation)* yang terdiri dari tiga komponen utama, yaitu *retriever*, *prompt template*, dan *language model (Mixtral-8x7B-Instruct-v0.1)*. Proses ini dimulai dari pengambilan bagian dokumen yang sesuai dengan pertanyaan, kemudian dikemas ke dalam format prompt, dan selanjutnya diproses oleh *language model* untuk menghasilkan respons. Kombinasi ini dirancang menggunakan *LangChain* sebagai kerangka kerja utama. Kode pembuatan RAG *Chain* dapat dilihat pada Gambar 2.

```
from langchain_core.output_parsers import StrOutputParser
from langchain_core.runnables import RunnablePassthrough


def format_docs(docs):
    return "\n\n".join(doc.page_content for doc in docs)

qa_chain = ( # pipeline
    {"context": retriever | format_docs, "question": RunnablePassthrough()}
    | custom_rag_prompt
    | llm
    | StrOutputParser()
)
```

Gambar 2. Kode RAG Chain

4.7 Sistem Menerima Pertanyaan dan Memberikan Jawaban

Pada bagian ini peneliti menggunakan dua perancangan untuk perbedaan hasil dalam menjawab. Pertama, dengan *language model* tanpa dokumen eksternal dan kedua dengan *retriever* yang hanya mengembalikan dokumen yang relevan.

a. *Language Model* Berbasis Konteks Dokumen *(Retriever* + LLM*)*

Sistem ini dirancang untuk menjawab pertanyaan pengguna dengan menggabungkan *retriever*, *prompt*, dan *language model (Mixtral-8x7B)*. Proses dimulai dengan pengguna mengajukan pertanyaan, contohnya: **"Kapan Anak Harus Mendapatkan Vaksin Hepatitis?"**. Pertanyaan ini digunakan oleh *retriever* untuk mencari dokumen yang relevan dari *vector database*. Dokumen yang ditemukan kemudian diformat dan dimasukkan ke dalam kerangka *prompt template*. Bahwa jawaban yang dihasilkan oleh *language* model berbasis pada konteks aktual dari dokumen. Prompt yang sudah terisi kemudian dikirimkan ke *LLM* untuk menghasilkan jawaban yang singkat, akurat, dan sesuai instruksi *template*. Terlihat pada Tabel 3, untuk contoh *input template* dan potongan konteks.

Tabel 3: Hasil *Retriever* + LLM

Hasil *Retriever* + LLM
Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer. Use three sentences maximum and keep the answer as concise as possible. Always say "thanks for asking!" at the end of the answer. … ... Vaksinasi hepatitis adalah bentuk perlindungan paling efektif untuk mencegah infeksi virus hepatitis. Anak perlu mendapat vaksin hepatitis A dan B untuk melindunginya dari infeksi virus tersebut. Jadwal vaksinasi hepatitis anak perlu diberikan secara lengkap agar perlindungannya optimal ... … QuestioKapan Anak Harus Mendapatkan Vaksin Hepatitis Helpful Answer:

b. *Retriever* untuk Mengembalikan Dokumen yang Relevan *(Retriever Only)*

Pada tahap ini, sistem hanya menggunakan *retriever* untuk menemukan dokumen yang relevan tanpa melalui *language model*. Ketika pengguna mengajukan pertanyaan, sistem akan mencari dan mengembalikan potongan teks yang paling relevan dari basis dokumen menggunakan vektor embedding.

Tabel 4: Hasil *Retriever Only*

Hasil *Retriever Only*
sebaiknya anak menerima 3 dosis vaksin hepatitis B saja. Bila saat bayi belum menerima vaksin hepatitis B, ketika remaja ia wajib mendapatkan vaksin. Pemberian vaksinasi hepatitis B wajib pada anak-anak dan remaja yang berusia kurang dari 19 tahun, mengingat infeksi virus bisa menyerang kapan saja. Terlebih jika kelompok ini tinggal di lingkungan atau negara endemik hepatitis B. Jadwal vaksin hepatitis untuk anak ini sebanyak 3 sampai 4 kali dengan masing-masing dosis berkisar antara 5–20 mg atau setara 0,5–1 ml. Dosis dan jadwal vaksinasi sangat bergantung dengan jenis vaksinasi hepatitis B yang digunakan. Untuk memastikannya, sebaiknya konsultasikan pada petugas kesehatan atau dokter yang bertugas secara langsung. Terkadang vaksin hepatitis B juga bergabung dengan vaksin hepatitis A, maka aturan pemberian vaksin dan dosisnya juga berbeda. Perlindungan vaksin hepatitis ini dapat bertahan ...

4.8 Pengujian

Sebagai bagian dari proses evaluasi terhadap kualitas jawaban yang dihasilkan oleh sistem, peneliti menggunakan metode *BERTScore* untuk mengukur tingkat kesamaan semantik antara jawaban model dan jawaban referensi. Evaluasi ini dilakukan dengan membandingkan jawaban sistem terhadap contoh pertanyaan **"Kapan Anak Harus Mendapatkan Vaksin Hepatitis?"** dengan jawaban ideal yang telah ditentukan. Berdasarkan hasil perhitungan menggunakan pustaka *bert_score*, diperoleh nilai *Precision* sebesar **0.8470%**, *Recall* sebesar **0.8026%**, dan *F1-Score* sebesar **0.8242**%. Nilai-nilai ini menunjukkan bahwa jawaban yang dihasilkan oleh model memiliki kemiripan semantik yang cukup tinggi terhadap referensi, meskipun terdapat sedikit perbedaan dalam struktur kalimat atau penggunaan kata.

5 Analisa Hasil

Pada tahap analisis, sistem *Retrieval-Augmented Generation* (RAG) yang dibangun menunjukkan kinerja yang baik dalam menjawab pertanyaan berbasis dokumen. Dua pendekatan dibandingkan, yaitu hanya menggunakan *retriever* dan kombinasi *retriever* dengan *language model (Mixtral-8x7B)*. Hasilnya, pendekatan RAG (*retriever* + LLM) mampu menghasilkan jawaban yang lebih ringkas, relevan, dan mudah dipahami dibandingkan potongan teks mentah dari *retriever* saja. Evaluasi menggunakan *BERTScore* menunjukkan tingkat kesamaan semantik yang cukup tinggi dengan nilai *Precision* sebesar 0.8470%, *Recall* 0.8026%, dan *F1-score* 0.8242%, yang menunjukkan bahwa jawaban sistem sudah cukup akurat terhadap jawaban referensi. *Template prompt* yang digunakan juga efektif dalam membatasi jawaban berdasarkan konteks dokumen dan mencegah halusinasi. Meskipun sistem memiliki kelebihan dalam efisiensi dan relevansi jawaban, keterbatasannya terletak pada ketergantungan terhadap kelengkapan dokumen dan kesulitan dalam menangani pertanyaan yang lebih kompleks. Secara keseluruhan, sistem ini berhasil menjawab pertanyaan dengan akurat dan sesuai konteks pada domain parenting.

6 Kesimpulan dan Saran

6.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan model *Mixtral-8x7B* dalam skema *Retrieval-Augmented Generation* (RAG) terbukti efektif dalam mengolah dan memahami data parenting. Sistem mampu menjawab pertanyaan pengguna secara relevan dan kontekstual berdasarkan dokumen yang tersedia, seperti pada contoh pertanyaan "Kapan Anak Harus Mendapatkan Vaksin Hepatitis?". Evaluasi menggunakan *BERTScore* menunjukkan nilai *F1-Score* sebesar 0.8242%, yang mengindikasikan bahwa jawaban sistem memiliki kemiripan semantik yang tinggi dengan jawaban referensi. Penggunaan data parenting dari Hello Sehat yang relevan dengan konteks Indonesia memperkuat kontribusi sistem terhadap pemahaman informasi pola asuh secara lokal. Dengan demikian, penelitian ini berhasil mencapai tujuannya, yaitu mengevaluasi efektivitas *Mixtral-8x7B* dalam menganalisis dokumen parenting secara efisien, sekaligus menunjukkan bahwa gabungan model *Mixtral* dengan metode RAG dapat menjadi solusi potensial dalam pengembangan sistem tanya jawab berbasis dokumen di ranah kesehatan dan keluarga.

6.2 Saran

Untuk pengembangan ke depan, sistem dapat ditingkatkan dengan menambahkan fitur *multi-turn* QA agar mampu menangani dialog bertahap. Kualitas dokumen sumber juga perlu ditingkatkan agar hasil pencarian lebih informatif dan komprehensif. Selain itu, penggunaan model yang telah di-*fine-tune* khusus pada topik parenting berpotensi meningkatkan kualitas jawaban. Terakhir, integrasi sistem ke dalam antarmuka berbasis chatbot atau aplikasi web akan memudahkan pengguna akhir dalam mengakses informasi parenting secara praktis dan cepat.

7 Referensi

[1] "These new estimates supersede former analyses and results published by UNICEF, WHO and the World Bank Group".

[2] D. Rokhanawati, H. Salimo, T. R. Andayani, and M. Hakimi, "The Effect of Parenting Peer Education Interventions for Young Mothers on the Growth and Development of Children under Five," *Children*, vol. 10, no. 2, Feb. 2023, doi: 10.3390/children10020338.

[3] L. Lasmini, B. Septiani, S. Aisyah, E. Selvia, and Y. F. Putri, "KONSEP DAN TAHAPAN PEMBENTUKAN PROGRAM PARENTING: KONSEP DAN TAHAPAN PEMBENTUKAN PROGRAM PARENTING," *Jurnal Multidisipliner Kapalamada*, vol. 1, no. 02, pp. 275–280, Jun. 2022, doi: 10.62668/kapalamada.v1i02.184.

[4] N. I. Putri, Y. Candrasari, P. Studi, and I. Komunikasi, "PESAN EDUKASI POSITIVE DISCIPLINE PARENTING PADA AKUN INSTAGRAM @GOODENOUGHPARENTS.ID," *JUITIK*, vol. 2, no. 2, 2022, [Online]. Available: http://journal.sinov.id/index.php/juitik/indexHalamanUTAMAJurnal:https://journal.sinov.id/index.php

[5] A. Titalessy and R. Y. Endang Kusumiati, "Dampak Perceraian Orang Tua Terhadap Perkembangan Sosial-Emosi Remaja," *Jurnal Ilmiah Bimbingan Konseling Undiksha*, vol. 12, no. 3, Nov. 2021, doi: 10.23887/jibk.v12i3.38582.

[6] A. A. Saleh and M. N. Hamang, "PENGASUHAN DISIPLIN POSITIF ISLAMI SEBAGAI UPAYA PENURUNAN KEKERASAN TERHADAP ANAK DI KABUPATEN SIDRAP".

[7] I. N. S. Paliwahet, I. M. Sukarsa, and I. K. Gede Darma Putra, “Pencarian Informasi Wisata Daerah Bali Menggunakan Teknologi Chatbot,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, p. 144, 2017, doi: 10.24843/lkjiti.2017.v08.i03.p01.

[8] M. R. Anam, A. S. Akbar, and H. Saputro, “QnA Chatbot with Mistral 7B and RAG method: Traffic Law Case Study,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 15, no. 03, p. 207, Jan. 2025, doi: 10.24843/lkjiti.2024.v15.i03.p06.

[9] A. Q. Jiang *et al.*, “Mixtral of Experts,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.04088

[10] untuk Penguatan Kesehatan dan Pemulihan Ekonomi Nasional Fitri Andri Astuti, “Pemanfaatan Teknologi Artificial Intelligence,” 2021.

[11] A. L. Maitri and J. Sutopo, “Rancang Bangun Chatbot sebagai Pusat Informasi Lembaga Kursus dan Pelatihan Menggunakan Pendekatan Natural Language Processing,” *University of Technology Yogyakarta*, 2019.

[12] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, Accessed: Jul. 09, 2025. [Online]. Available: https://arxiv.org/pdf/2005.11401

[13] R. Islam, *Retrieval-Augmented Generation (RAG): Empowering Large Language Models (LLMs)*. Dr. Ray Islam (Mohammad Rubyet Islam), 2023. [Online]. Available: https://books.google.co.id/books?id=5xRm0AEACAAJ

[14] L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O’Reilly Media, 2022. [Online]. Available: https://books.google.co.id/books?id=pNBpzwEACAAJ

[15] S. Elysia and Herianto, “Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah,” *Journal TIFDA (Technology Information and Data Analytic)*, vol. 1, no. 2, pp. 52–58, 2024, doi: 10.70491/tifda.v1i2.52.

[16] A. T. U. BR. Lubis, N. S. Harahap, S. Agustian, M. Irsyad, and I. Afrianty, “Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan),” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 955–964, May 2024, doi: 10.57152/malcom.v4i3.1378.