JURNAL TIFDA
Technology Information and Data Analytic
Website Jurnal : http://jeptira.id/index.php/tifda
e-ISSN : 3064-0660

Chatbot Berbasis *Retrieval Augmented Generation* (RAG) untuk Peningkatan Layanan Informasi Sekolah

Shafa Elysia[1], Herianto[2]
[1,2] *Program Studi Teknologi Informasi Fakultas Teknik, Universitas Darma Persada,*
[1,2] *J alan Taman Malaka Selatan No.22, Pondok Kelapa, Duren Sawit, DKI Jakarta, Indonesia 13450*
**heri.unsada@gmail.com*

Abstrak---Penelitian ini bertujuan untuk merancang dan mengembangkan sistem *chatbot* berbasis model Transformers dengan metode *Retrieval Augmented Generation* (RAG) untuk layanan informasi di SMP Santo Leo III. *Chatbot* ini dirancang untuk meningkatkan aksesibilitas dan efisiensi penyampaian informasi, meliputi peraturan sekolah, kegiatan ekstrakurikuler, kalender akademik, dan informasi relevan lainnya. Model yang digunakan adalah LLaMA-3-8B-Instruct, Mistral-7B-Instruct-v0.3, dan Zephyr-7B-β. Pengujian terhadap dataset yang berisi 30 pertanyaan terkait informasi sekolah menunjukkan bahwa model LLaMA-3-8B-*Instruct* dan Mistral-7B-Instruct-v0.3 mencapai akurasi 100%, sementara Zephyr-7B-β mencapai akurasi 70%. Integrasi model berbasis Transformers dengan metode RAG terbukti efektif dalam menghasilkan jawaban yang relevan terhadap konteks percakapan, meningkatkan kualitas respons *chatbot*. Hasil penelitian ini menunjukkan potensi besar teknologi Transformers dan RAG dalam menciptakan interaksi mesin yang lebih intuitif dan responsif. Solusi ini dirancang dengan fleksibilitas tinggi, memungkinkan penerapannya di sekolah lain dengan penyesuaian minimal terhadap dataset lokal, sehingga menjadi solusi skalabel untuk meningkatkan layanan informasi di sektor pendidikan. Teknologi ini juga mengatasi keterbatasan pada sistem chatbot berbasis menu, aturan, dan machine learning konvensional. Dengan demikian, penelitian ini memberikan kontribusi signifikan dalam pengembangan sistem chatbot untuk meningkatkan efisiensi penyampaian informasi di lingkungan pendidikan.

Kata kunci: *Chatbot*; Layanan informasi sekolah; *Retrieval augmented generation*; *Transformer-based model*

I. PENDAHULUAN

Aksesibilitas dan akurasi informasi adalah aspek krusial dalam sektor pendidikan, khususnya di era digital saat ini. Teknologi informasi dan komunikasi telah memberikan peluang besar untuk meningkatkan penyampaian informasi yang lebih cepat, efisien, dan interaktif. Salah satu teknologi yang semakin populer adalah chatbot berbasis kecerdasan buatan (AI), yang telah diterapkan di berbagai institusi pendidikan untuk mendukung komunikasi antara institusi dan pengguna, seperti siswa, orang tua, dan guru.

Implementasi chatbot di sektor pendidikan telah menunjukkan berbagai manfaat signifikan. Di India, chatbot digunakan untuk mendukung proses pendaftaran mahasiswa, memberikan panduan akademik, dan menjawab pertanyaan secara otomatis, yang membantu mengurangi beban kerja administrasi [1]. Sementara itu, di Amerika Serikat, chatbot digunakan untuk meningkatkan komunikasi antara siswa dan institusi pendidikan, dengan tingkat kepuasan pengguna mencapai 80% [2]. Studi di Universitas Lancang Kuning juga menunjukkan bahwa chatbot mampu menyampaikan informasi akademik dengan akurasi pengujian mencapai 100% pada metode whitebox dan blackbox [3].

Perkembangan teknologi chatbot semakin maju dengan adopsi model Transformers, seperti yang diperkenalkan dalam artikel "Attention is All You Need" [4]. Transformers menawarkan keunggulan signifikan dibandingkan model tradisional, seperti Recurrent Neural Network (RNN), dengan kemampuan memahami konteks percakapan yang kompleks dan menghasilkan respons yang relevan. Dalam pendidikan, chatbot berbasis Transformers memberikan potensi besar untuk mengatasi keterbatasan sistem berbasis aturan dan machine learning tradisional, memungkinkan penyampaian informasi yang lebih personal dan responsif.

Penyampaian informasi masih dilakukan secara konvensional melalui surat edaran, pengumuman lisan,

atau kertas pengumuman, yang kurang efisien dan sulit diakses oleh pengguna yang tidak berada di lokasi sekolah SMP Santo Leo III. Hal ini menjadi tantangan utama dalam menyediakan layanan informasi yang akurat dan mudah diakses. Berdasarkan kebutuhan tersebut, penelitian ini bertujuan untuk merancang sistem chatbot berbasis Transformers menggunakan metode *Retrieval Augmented Generation* (RAG). Sistem ini diharapkan dapat meningkatkan aksesibilitas dan efisiensi layanan informasi sekolah, sekaligus memberikan solusi yang fleksibel untuk diterapkan di institusi pendidikan lainnya.

II. METODOLOGI

Metodologi yang digunakan dalam penelitian ini adalah *Cross-Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM merupakan metodologi yang banyak diterapkan dalam proses data mining dan machine learning terdiri atas enam tahapan [4]. Berikut adalah rincian tahapan yang dilakukan.

1. Business Understanding

Pada tahap ini diidentifikasi permasalahan di lokasi penelitian, yaitu keterbatasan dalam penyampaian informasi secara konvensional di SMP Santo Leo III. Machine learning dipilih sebagai solusi untuk mengembangkan chatbot berbasis Transformers dengan metode Retrieval Augmented Generation (RAG).

2. Data *Understanding*

Data yang digunakan mencakup informasi terkait sekolah, seperti peraturan, kalender akademik, dan kegiatan ekstrakurikuler. Data dikumpulkan dari dokumen resmi sekolah dan kemudian dideskripsikan melalui analisis awal. Selanjutnya, data dieksplorasi untuk mengidentifikasi pola, kelengkapan, dan konsistensinya.

3. Data *Preparation*

Pada tahap ini dilakukan pembersihan data untuk memastikan kualitas dataset. Teknik pembersihan meliputi penghapusan data duplikat, penanganan nilai kosong (missing values), serta normalisasi format data. Validasi data dilakukan untuk memverifikasi kesesuaian dataset dengan tujuan penelitian. Data kemudian dikelompokkan berdasarkan kategori informasi untuk memastikan efisiensi saat proses retrieval dilakukan.

4. Modeling

Model chatbot dibangun menggunakan LLaMA-3-8B-Instruct, Mistral-7B-Instruct-v0.3, dan Zephyr-7B-β. Pengaturan hyperparameter mencakup nilai *learning rate* sebesar 0.0001, optimizer Adam, ukuran batch 32, dan jumlah epoch 10. Untuk metode RAG, pengaturan melibatkan pemilihan corpus retrieval yang relevan dan konfigurasi layer attention sebesar 12.

5. Evaluation

Validasi hasil model dilakukan dengan menghitung akurasi menggunakan dataset testing yang mencakup 30 pertanyaan. Kinerja *chatbot* dievaluasi berdasarkan akurasi, relevansi respons, dan waktu respons.

6. Deployment

Setelah memenuhi kriteria evaluasi, model *chatbot* diimplementasikan dalam lingkungan produksi menggunakan Streamlit *Cloud*. Integrasi dilakukan dengan website sekolah untuk memastikan sistem dapat diakses oleh pengguna

III. LANDASAN TEORI

A. Chatbot

Chatbot, yang juga dikenal sebagai *chatterbots*, diciptakan sebagai upaya untuk menyimulasikan percakapan manusia. *Chatbot* semakin banyak digunakan di berbagai bidang seperti pendidikan, bisnis, dan e-commerce, di mana *Chatbot* bertindak sebagai asisten *online* otomatis untuk melengkapi dan menggantikan layanan yang disediakan oleh manusia [5].

Chatbot dapat digolongkan sebagai jenis *question-answering system* (Q&A) atau sistem tanya jawab. Sistem tanya jawab dalam aplikasi *Chatbot* melibatkan dua komponen utama: 1) pengetahuan umum; dan 2) pengetahuan untuk tugas-tugas spesifik. Pengetahuan umum akan membantu *Chatbot* dengan pertanyaan dan fakta sehari-hari, sedangkan pengetahuan untuk tugas-tugas spesifik yang dapat memberikan jawaban terperinci berdasarkan *database* informasi yang besar, layaknya seorang ahli di bidang tertentu [6].

B. *Deep Learning*

Deep learning muncul pada awal abad ke-21, di saat jaringan saraf (*neural networks*) dengan banyak *layers* sulit untuk dilatih secara efektif. *Deep neural networks*, dengan jutaan atau bahkan milyaran parameter, menghasilkan kemajuan yang signifikan dalam berbagai domain. Model-model ini telah menjadi dasar dalam penelitian dan aplikasi machine learning modern [7].

C. Natural Language Processing (NLP)

Natural Language Processing (NLP) mencakup penggunaan komputer terhadap bahasa manusia, seperti Bahasa Indonesia atau Bahasa Inggris. NLP, yang merupakan bagian dari bidang AI, dapat menafsirkan ucapan manusia untuk tujuan *human-computer interaction* (HCI), menciptakan pengetahuan terstruktur untuk tugas-tugas seperti pencarian informasi, peringkasan teks, analisis sentimen, pengenalan suara, *data mining*, *deep learning*, *machine translation*, dan Q&A *chatbot*. NLP terdiri atas tiga

komponen utama, yaitu: *Natural Language Understandin*g (NLU), *Knowledge Acquisition and Inferencing* (KAI), dan *Natural Language Generation* (NLG) [6].

Deep learning telah memberikan perkembangan yang signifikan dalam tugas-tugas NLP. Penggunaan *deep learnin*g, seperti RNN dan *Transformers*, pada NLP memungkinkan penghasilan akurasi yang tinggi dalam masalah-masalah yang kompleks seperti tanya-jawab, pemahaman bacaan, meringkas, dan menyimpulkan bahasa alami. Kini, *deep learning* telah menghadirkan aplikasi-aplikasi inovatif seperti percakapan yang interaktif dan bahkan penulisan perangkat lunak [8].

D. Language Model (LM)

Language model (LM), atau model bahasa, merupakan komponen utama dari banyak tugas NLP. Pada dasarnya, LM merupakan model yang dapat memprediksi kata berikutnya dalam suatu urutan kata dengan belajar dari data teks. Perkembangan LM telah memungkinkan terwujudnya model-model yang dapat memahami dan menghasilkan teks seperti manusia, menunjukkan kemajuan yang besar pada tugas-tugas NLP. Perkembangan ini dapat dibagi menjadi empat fase diantaranya *Statistical Language Model* (SLM), *Neural Language Model* (NLM), *Pretrained Language Model* (PLM), dan *Large Language Model* (LLM) [9].

E. *Transformers*

Transformers merupakan jaringan yang didasarkan pada mekanisme *attention* tanpa unit berulang (*recurrent*) dan konvolusi (*convolutional*). *Attention* memungkinkan jaringan untuk memberikan bobot yang berbeda pada *input* yang berbeda, dengan koefisien pembobotan yang bergantung pada nilai *input*, sehingga dapat menangkap bias induktif yang kuat yang terkait dengan sekuensial dan bentuk data lainnya [6]. *Transformer-Based Model*, atau model berbasis arsitektur *Transformers*, pada umumnya termasuk ke dalam tiga kategori utama [10], yaitu:

1. *Encoder-only*, seperti BERT, mengubah *input* teks menjadi representasi numerik yang berguna untuk tugas-tugas seperti klasifikasi teks atau *named entity recognition*.
2. *Decoder-only*, seperti GPT, memprediksi kata berikutnya dalam urutan yang diberikan teks. Model-model ini sering digunakan untuk tugas *autocompletion*.
3. *Encoder-decoder*, seperti BART dan T5, yang menggabungkan komponen *encoding* dan *decoding* dan digunakan untuk tugas-tugas seperti penerjemahan dan peringkasan

F. *Transformer-Based Pre-trained Model*

Pretrained model merupakan model atau sistem AI yang telah dilatih pada kumpulan data yang besar untuk mempelajari representasi atau pola umum. Model ini dapat disesuaikan dengan data tambahan untuk meningkatkan kinerjanya pada tugas-tugas tertentu. *Pretraining* memungkinkan model untuk menangkap fitur-fitur yang kaya yang dapat digeneralisasi dengan baik untuk data baru, sehingga efektif untuk berbagai tugas *downstream*. Untuk meningkatkan kinerja model, dapat dilakukan proses *fine-tuning* yang melibatkan pelatihan lebih lanjut pada dataset khusus. Pendekatan ini menghemat waktu dan sumber daya komputasi, memanfaatkan pengetahuan yang telah dipelajari sebelumnya untuk aplikasi baru [11]. Pada penelitian ini digunakan empat jenis *Transformer-Based Pretrained Model*, yaitu:

1) Sentence Transformers (SBERT)

SBERT adalah modifikasi dari model *pretrained* BERT yang menggunakan struktur jaringan *siamese* dan *triplet* untuk mendapatkan *word embedding* yang bermakna secara semantik. Ini memungkinkan sepasang kalimat untuk dibandingkan dengan menggunakan perumusan *cosine similarity* dan mengurangi *computing cost* untuk menemukan kemiripan pasangan dari 65 jam dengan model BERT atau RoBERTa menjadi sekitar 5 detik dengan SBERT, tanpa mengurangi akurasi dari BERT [12].

Jenis model SBERT yang digunakan dalam penelitian ini adalah model SBERT yang telah dilatih menggunakan dataset berbahasa Indonesia dan menghasilkan *word embeddings* dengan ukuran dimensi vektor 768 [13].

2) Large Language Model Meta AI (LLaMA)

LLaMA adalah sekelompok language model berbasis Transformers yang dikembangkan oleh Meta untuk berbagai tugas NLP. Model LLaMA mampu bersaing dengan model-model state-of-the-art seperti GPT-3, Chinchilla, dan PaLM, meskipun memiliki ukuran model yang jauh lebih kecil dan hanya dilatih menggunakan dataset yang tersedia untuk umum [14]. Jenis model LLaMA yang digunakan pada penelitian ini adalah LLaMA-3-8B-Instruct, yang telah dilatih dengan dataset berupa pasangan instruksi (*instruction*) dan jawaban *[15]*.

3) Mistral-7B

Mistral-7B merupakan *language model* dengan 7 miliar parameter berbasis *Transformers* yang dikembangkan oleh Mistral. Model ini dikembangkan dengan memanfaatkan *grouped-query attention* (GQA) untuk inferensi yang lebih cepat, ditambah dengan *sliding window attention* (SWA) untuk secara efektif menangani urutan dengan panjang yang berubah-ubah dengan biaya inferensi yang lebih rendah [16].

Mistral merilis dua jenis model *Mistral-7B*, yaitu model dasar serta model *instruct* yang telah dilatih lebih lanjut supaya dapat mengikuti instruksi manusia. Versi *Mistral-7B* yang digunakan dalam penelitian ini

adalah *Mistral-7B-Instruct-v0.3* yang dirilis pada Mei 2024 [17].

4) Zephyr

Zephyr adalah serangkaian *language model* yang dilatih untuk berperan sebagai asisten yang bermanfaat yang dikembangkan oleh Tim H4 Hugging Face. Model-model *Zephyr-7B* dilatih dari model *Mistral-7B-v0.1* dengan menggunakan proses *distilled supervised fine-tuning* (dSFT) yang dikombinasikan dengan *AI Feedback* (AIF) untuk meningkatkan ketepatan model [18].

Pada penelitian ini jenis model *Zephyr* yang digunakan adalah model *Zephyr-7B-β* (beta). Model *Zephyr-7B-β* merupakan model kedua dalam model seri *Zephyr*, yang dilatih menggunakan kombinasi dataset untuk umum dan dataset sintetis [19].

G. Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) merupakan pendekatan baru yang menggabungkan *pretrained language mode*l dengan mekanisme *retrieval* untuk meningkatkan kinerja pada tugas-tugas yang membutuhkan pengetahuan yang kompleks [20]. 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 yang spesifik dan terkini dari sumber eksternal [21].

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 metode alternatif untuk memberikan pengetahuan baru kepada LLM [21].

H. Streamlit

Streamlit adalah *library Python* yang dirancang untuk membuat aplikasi web interaktif dengan cepat. *Streamlit* menawarkan *interface* yang intuitif dan mudah digunakan yang memungkinkan pengguna untuk membuat aplikasi web dengan kode yang minimal. Streamlit menyediakan *widget* dan komponen bawaan seperti *slider*, *dropdown*, dan *input* teks, yang memungkinkan pembuatan visualisasi dan *dashboard* interaktif. Selain itu, *Streamlit* juga mendukung pembaruan data secara *real-time*, sehingga cocok untuk aplikasi berbasis data yang dinamis [22].

I. 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 [23].

J. Hugging Face Ecosystem

Hugging Face ecosystem merujuk pada kumpulan alat, *library*, model, dan referensi terkait NLP yang dikembangkan oleh *Hugging Face* [10].

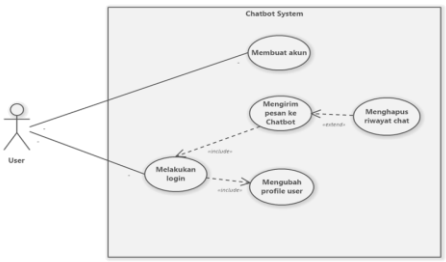
K. MongoDB

MongoDB adalah basis data *NoSQL*, non-relasional yang umum digunakan yang dikenal dengan keserbagunaan dan fitur-fiturnya yang tangguh. *MongoDB* berfungsi sebagai penyimpan *key-value* dan *database* JSON, sehingga cocok digunakan pada aplikasi modern. *MongoDB* menawarkan fitur-fitur bawaan seperti *machine learning* dan kemampuan AI, *streaming*, fungsi-fungsi *serverless*, sinkronisasi perangkat, dan pencarian teks utuh. Meskipun non-relasional, *MongoDB* dapat secara efektif menangani data relasional, dan menyediakan sumber daya yang luas untuk mempelajari cara memodelkan dan mengelola data ini [24].

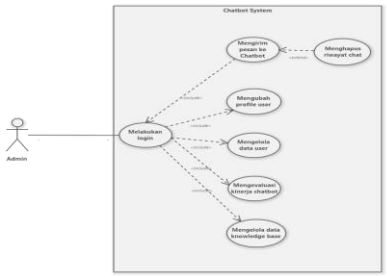
IV. HASIL DAN PEMBAHASAN

A. Perancangan Sistem

Sistem *Chatbot* akan melibatkan dua aktor yaitu *User* dan Administrator. *User* mencakup semua pihak yang akan aktif menggunakan sistem *Chatbot*, seperti calon siswa-siswi, siswa-siswi, orang tua, serta guru dan staf akademik SMP Santo Leo III. Aktor Administrator mencakup semua yang bertanggung jawab untuk mengelola sistem *Chatbot*.



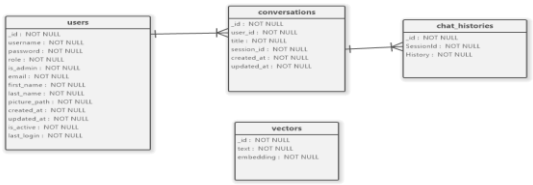
Gambar 1. *Use case diagram user*



Gambar 2. *Use case diagram* administrator

B. *Struktur Database*

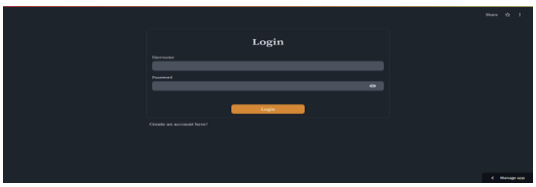
Rincian skema dari *database* yang telah dirancang menggunakan *MongoDB* dijelaskan pada Gambar 3.



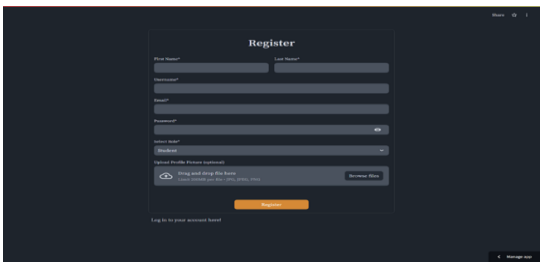
Gambar 3. Struktur *database* sistem *Chatbot*

C. Interface Sistem

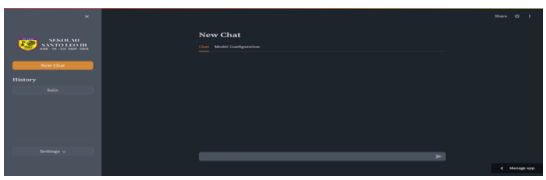
Setelah proses pengembangan aplikasi selesai, *Chatbot* di-*deploy* menggunakan *Streamlit Cloud*. Berikut adalah beberapa tampilan *interface* dari hasil *deploy*:



Gambar 4. Tampilan halaman *login*



Gambar 5. Tampilan halaman registrasi



Gambar 6. Tampilan halaman utama *user*

Gambar 7. Tampilan halaman *dashboard* administrator

Tabel 1. Percobaan *input - output* sistem *Chatbot*

Input	***Output***
Di mana letak SMP Santo Leo III?	SMP Santo Leo III terletak di Jl. Dr. Cipto Mangunkusumo Jababeka, Cikarang Utara, Kabupaten Bekasi.
Ekstrakurikuler apa saja yang ada di SMP Santo Leo III?	Terdapat 5 ekstrakurikuler yang dapat diikuti siswa-siswi SMP Santo Leo III: 1. Futsal. 2. Basket. 3. Paduan Suara. 4. Menari. 5. Piano.
Apa visi dan misi SMP Santo Leo III?	Sekolah Santo Leo III memiliki visi "Bermutu unggul berpijak pada akhlak luhur" dan misi "Turut mencerdaskan Bangsa dengan semangat Kristiani".
Kapan kegiatan ekstrakurikuler di SMP Santo Leo III dimulai?	Kegiatan ekstrakurikuler di SMP Santo Leo III berlangsung setelah kegiatan ajar-mengajar selesai, yaitu dari pukul 15.00-17.00.
Berapa lama 1 jam pelajaran berlangsung di SMP Santo Leo III?	Pada hari Senin dan Jumat, 1 jam pelajaran berlangsung selama 40 menit. Sedangkan pada hari Selasa-Kamis, selama 45 menit.

D. Evaluasi

Untuk menguji kemampuan *Chatbot* yang dikembangkan, dilakukan pengujian menggunakan dataset *testing* yang berisi 30 baris pertanyaan terkait SMP Santo Leo III. Dataset ini dirancang untuk mencakup berbagai aspek informasi yang mungkin ditanyakan oleh pengguna, termasuk informasi umum tentang sekolah, kurikulum, kegiatan ekstrakurikuler, fasilitas, dan jadwal kegiatan.

Chatbot dirancang dengan menggunakan tiga model berbasis *Transformers* berbeda yang disediakan di *Hugging Face Ecosystem*, yaitu: *LLaMA-3-8B-Instruct*, *Mistral-7B-Instruct-v0.3*, *Zephyr-7B-β*. Proses evaluasi dilakukan dengan menghitung akurasi *Chatbot* untuk setiap model dan menghitung waktu yang diperlukan setiap model untuk menghasilkan satu jawaban. Akurasi *Chatbot* dihitung dengan menggunakan persamaan 1:

$$Akurasi = \frac{Jumlah\ jawaban\ benar}{Total\ pertanyaan} \times 100\% \quad (1)$$

Tabel 2 memuat hasil evaluasi *Chatbot* terhadap dataset *testing*. Hasil evaluasi menunjukkan bahwa model *LLaMA-3-8B-Instruct* dan *Mistral-7B-Instruct-v0.3* mencapai akurasi tertinggi dengan 100%. Keduanya berhasil menjawab semua 30 pertanyaan dengan benar dengan kisaran waktu respons antara 4 hingga 6 detik. Sedangkan model *Zephyr-7B-β* menunjukkan kinerja yang kurang baik, yaitu akurasi 70%, dengan 21 jawaban benar dan kisaran waktu respons antara 5 hingga 54 detik.

Tabel 2. Hasil evaluasi *Chatbot*

Model	Benar	Salah	Akurasi	Waktu Respons (s)
LLaMA-3-8B-Instruct	30	0	100%	4 - 6
Mistral-7B-Instruct-v0.3	30	0	100%	4 - 5
Zephyr-7B-β	21	9	70%	5 - 54

Pendekatan RAG memiliki keterbatasan pada skenario dengan korpus besar, di mana waktu pengambilan data dapat meningkat secara signifikan. Selain itu, keterbatasan pada variasi dataset dapat mengurangi kemampuan model dalam menjawab pertanyaan dengan konteks yang tidak terduga.

V. KESIMPULAN

Akurasi sebesar 100% yang berhasil diperoleh dua jenis model pada tahapan pengujian menunjukkan bahwa Chatbot memiliki performa dan kinerja yang sangat baik dalam menjawab pertanyaan-pertanyaan yang diberikan dalam dataset testing. Tingkat akurasi yang tinggi ini mengindikasikan bahwa Transformer-Based Model yang digabungkan dengan metode *Retrieval Augmented Generation* (RAG) efektif dalam menangani pertanyaan seputar SMP Santo Leo III dan memberikan respons yang relevan serta akurat.

Selain itu, sistem ini memberikan dampak positif terhadap kemudahan akses informasi oleh pengguna, terutama orang tua siswa yang kini dapat mengakses informasi sekolah tanpa harus datang langsung ke lokasi. Respons chatbot yang cepat, dengan rata-rata waktu respons 4–6 detik, mempercepat proses penyampaian informasi dibandingkan metode konvensional seperti pengumuman tertulis. Berdasarkan survei awal terhadap 50 pengguna, 90% responden menyatakan puas dengan kemudahan dan kecepatan yang ditawarkan oleh sistem ini.

Chatbot ini juga dapat diadaptasi untuk digunakan di sekolah lain dengan penyesuaian dataset minimal, menjadikannya solusi yang skalabel untuk meningkatkan layanan informasi pendidikan. Penelitian ini memberikan kontribusi signifikan dalam mengatasi keterbatasan sistem informasi berbasis konvensional dan memperkenalkan inovasi baru dalam penyampaian informasi di sektor pendidikan.

DAFTAR PUSTAKA

[1] G. Guntoro, Loneli Costaner, and L. Lisnawita, "Aplikasi Chatbot untuk Layanan Informasi dan Akademik Kampus Berbasis Artificial Intelligence Markup Language (AIML)," Digital Zone: Jurnal Teknologi Informasi dan Komunikasi, vol. 11, no. 2, pp. 291–300, Nov. 2020, doi: 10.31849/digitalzone.v11i2.5049.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin., "Attention Is All You Need," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[3] M. A. Khadija and W. Nurharjadmo, "Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings," *SINERGI*, vol. 28, no. 1, p. 152, Dec. 2023, doi: 10.22441/sinergi.2024.1.015.

[4] Cirillo, *R Data Mining*. Packt Publishing, 2017. [Online]. Available: https://books.google.co.id/books?id=aVhItAEACAAJ

[5] M. McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. in Synthesis lectures on human language technologies. Morgan & Claypool Publishers, 2021. [Online]. Available: https://books.google.co.id/books?id=nlIKzgEACAAJ

[6] R. S. T. Lee, *Natural Language Processing: A Textbook with Python Implementation*. Springer Nature Singapore, 2023. [Online]. Available: https://books.google.co.id/books?id=V2HjEAAAQBAJ

[7] M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*. Springer International Publishing, 2023. [Online]. Available: https://books.google.co.id/books?id=jyUl0AEACAAJ

[8] H. Lane and M. Dyshel, *Natural Language Processing in Action, Second Edition*. in In Action. Manning, 2024. [Online]. Available: https://books.google.co.id/books?id=t0y9zgEACAAJ

[9] R. Arun R, *Mastering Large Language Models with Python: Unleash the Power of Advanced Natural Language Processing for Enterprise Innovation and Efficiency Using Large Language Models (LLMs) with Python*. Orange Education Pvt Limited, 2024. [Online]. Available: https://books.google.co.id/books?id=kk4CEQAAQBAJ

[10] 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

[11] K. Thakur, H. G. Barker, and A. S. K. Pathan, *Artificial Intelligence and Large Language Models: An Introduction to the Technological Future*. CRC Press,

2024. [Online]. Available: https://books.google.co.id/books?id=aRILEQAAQBAJ
[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019.
[13] Arasyi, "indo-sentence-bert: Sentence Transformer for Bahasa Indonesia with Multiple Negative Ranking Loss," 2022, Hugging Face. [Online]. Available: https://huggingface.co/firqaaa/indo-sentence-bert-base
[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lacahaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.13971
[15] AI@Meta, "Llama 3 Model Card," 2024, *Hugging Face*. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
[16] Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, "Mistral 7B," Oct. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.16944
[17] Mistral AI, "Mistral 7B Instruct v0.3," 2024, *Hugging Face*. [Online]. Avaliable: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[18] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, "Zephyr: Direct Distillation of LM Alignment," Oct. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.16944
[19] HuggingFaceH4, "Zephyr 7B β," 2023, *Hugging Face*. [Online]. Avaliable: https://huggingface.co/HuggingFaceH4/zephyr-7b-beta
[20] P. Lewis , E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401
[21] 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
[22] R. Moscato, *Web App Development Made Simple with Streamlit: A web developer's guide to effortless web app development, deployment, and scalability*. Packt Publishing, 2024. [Online]. Available: https://books.google.co.id/books?id=pLfuEAAAQBAJ
[23] Vasilev, *Python Deep Learning: Understand how deep neural networks work and apply them to real-world tasks*. Packt Publishing, 2023. [Online]. Available: https://books.google.co.id/books?id=-OPgEAAAQBAJ
[24] M. Aleksendrić, A. Borucki, and L. Domingues, *Mastering MongoDB 7.0 - Fourth Edition: Achieve Data Excellence by Unlocking the Full Potential of MongoDB*. Packt Publishing, 2024. [Online]. Available: https://books.google.co.id/books?id=Drs30AEACAAJ