BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Pada tabel 2.1 terdapat informasi mengenai penelitian terdahulu yang dijadikan sebagai pendukung dalam penelitian yang dilakukan. Dalam menganalisis penelitian-penelitian tersebut, ada beberapa aspek yang perlu ditentukan terlebih dahulu, yaitu permasalahan, *research gap*, dan hasil yang diperoleh. Setiap penelitian muncul sebagai respons terhadap permasalahan yang perlu dipecahkan. Pencarian permasalahan ini dapat ditemukan dalam bagian latar belakang dan rumusan masalah. Selanjutnya, akan dianalisis celah penelitian yang terdapat dalam jurnal, mulai dari analisis bagian abstrak, tujuan, metode, temuan, hingga kesimpulan. Dari analisis tersebut, dapat diidentifikasi teori yang belum ada atau kekurangan dalam hasil penelitian. Hasil penelitian juga dapat dievaluasi berdasarkan keberhasilan yang dicapai. Dalam penelitian ini, dilakukan analisis yang mengadopsi beberapa elemen dari penelitian sebelumnya untuk diterapkan dalam penelitian yang sedang berlangsung. Berikut adalah rincian dari masing-masing penelitian sebelumnya beserta elemen-elemen yang diadopsi.

Tabel 2.1 Penelitian Terdahulu

Penelitian Terdahulu		
1.	Nama Jurnal	arXiv
	Tahun Penerbitan	2023
	Judul	Mistal 7B
	Penulis	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed [23].
	Metode Penelitian	Penelitian melalukan ujicoba dengan metode *commonsense Reasoning (0 Shot), World Knowledge (5-shot), Reading Comprehension (0-shot), Math, Code,* dan *Popular aggregated result.* Penelitian ini juga mengujicobakan efektivitas terhadap *Instruction*

Penelitian Terdahulu		
		Finetuning terhadap model yang dikomparasi terhadap model lainnya.
	Hasil	Pada setiap metrik yang diujicobakan, Mistral 7B berhasil mendapatkan nilai terbaik dibandingkan model lainnya. Pada ujicoba prompt, Mistral 7B dapat memberikan hasil yang diperlukan dengan menjawab pertanyaan "*how to kill linux process*", sedangkan pada Llama 2 13B menolak untuk menjawab pertanyaan yang sama.
2.	Nama Jurnal	arXiv
	Tahun Penerbitan	2023
	Judul	InstructPTS: Instruction-Tuning LLMs for Product Title Summarization.
	Penulis	Besnik Fetahu Zhiyu Chen Oleg Rokhlenko Shervin Malmasi [26].
	Metode Penelitian	Penelitian menggunakan berbagai model LLM dengan menggunakan *supervised finetuning, Control Codes* dan *Instruction Tuning* dan dievaluasikan menggunakan BLEU dan ROUGE Score.
	Hasil	Hasil yang didapatkan pada model Flan -T5 dengan menggunakan *Intruction Tuning Approach* mendapatkan akurasi pada online deployment dengan 4 ribu sampel judul produk sebesar 96% telah memenuhi kriteria yang ditetapkan.
3.	Nama Jurnal	MDPI
	Tahun Penerbitan	2025
	Judul	Fine-Tuning Large Language Models for Ontology Engineering: A Comparative Analysis of GPT-4 and Mistral.
	Penulis	Dimitrios Doumanas, Andreas Soularidis, Dimitris Spiliotopoulos, Costas Vassilakis, and Konstantinos Kotis [27].
	Metode Penelitian	Penelitian menggunakan 15 Q&A yang disiapkan oleh GPT-4 dan ditanyakan kepada model untuk diketahui *true-positives, false-positives, false-positives, false negatives.*
	Hasil	Pada ujicoba pertama, GPT-4 mendapatkan nilai recall sebesar 11.25% hingga 16.25%, sedangkan Mistral 7B mendapatkan nilai recall sebesar 15.25%. Pada percobaan kedua, GPT-4 mengalamai kenaikan performa dengan mencapai nilai F1 Score 0.4210, dengan nilai recall 23.33%. pada Mistral 7B menunjukan performa yang lebih baik pada evaluasi kelas dan object property metrics dengan nilai F1 Score sebesar 0.5379 dan nilai recall 48.75% .

Penelitian Terdahulu		
4.	Nama Jurnal	RSNA Journal, Radiology
	Tahun Penerbitan	2025
	Judul	Assessing Completeness of Clinical Histories Accompanying Imaging Orders Using Adapted Open-Source and Closed-Source Large Language Models.
	Penulis	David B. Larson, Arogya Koirala, Lina Y. Cheuy, Magdalini Paschali, Dave Van Veen, Hye Sun Na, Matthew B. Petterson, Zhongnan Fang Akshay S. Chaudhari [28].
	Metode Penelitian	Model penelitian dilatih dengan data historis Kesehatan termasuk CT, MRI, US dan radiografis untuk diberikan perintah untuk mengekstrak elemen dari riwayat medis, apa, kapan, dan dimana dari permasalahan Riwayat klinis.
	Hasil	Dari dua model *open-source*, Mistral-7B mengungguli Llama 2–7B dalam menilai kelengkapan dan selanjutnya disesuaikan lebih lanjut. Baik Mistral-7B maupun GPT-4 Turbo menunjukkan hasil yang substansial dengan radiolog (rata-rata κ, 0,73 [95% CI: 0,67, 0,78] hingga 0,77 [95% CI: 0,71, 0,82]).
5.	Nama Jurnal	arXiv
	Tahun Penerbitan	2024
	Judul	Birbal: An efficient 7B instruct-model fine-tuned with curated datasets
	Penulis	Ashvini Kumar Jindal, Pawan Kumar Rajpoot, Ankur Parikh [29]
	Metode Penelitian	Penelitian ini menggunakan model Mistral-7B dengan ukuran 200K, 400K, dan 700K ukuran dataset dengan 3 epoch, 2 epoch, dan 1 epoch. Setiap model tersebut dilatih menggunakan data kurasi dan *fine tuning*.
	Hasil	Hasil mode yang telah dibuat dilakukan evaluasi melalui pertandingan untuk dibandingkan dengan 57 model lainnya. Hasilnya didapatkan model Mistral 200K mendapatkan nilai evaluasi tertinggi dari berbagai dataset dan metrik yang diujicobakan dengan nilai evaluasi 0.660.
6.	Nama Jurnal	JMIR
	Tahun Penerbitan	2024
	Judul	Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: An Evaluation of Strategies and Performance
	Penulis	Océane Dorémus, Dylan Russon, Benjamin Contrand, Ariel Guerra-Adames, Marta Avalos-Fernandez, Cédric Gil-Jardiné, Emmanuel Lagarde PhD [30].

Penelitian Terdahulu		
	Metode Penerbitan	Penelitian menggunakan dataset berupa 425 Ribu catatan kesehatan yang akan memvalidasi data berupa referensi yang telah disiapkan
	Hasil	Model generatif Mistral 7B menunjukkan kinerja tertinggi dengan skor F1 keseluruhan sebesar 0,9673 (dibandingkan dengan 0,8750 untuk Llama 2 dan 0,8686 untuk Mixtral 8x7B). Pada tingkat catatan klinis, model yang sama mencapai recall keseluruhan sebesar 0,9326 (dibandingkan dengan 0,6888 untuk Llama 2 dan 0,6417 untuk Mixtral 8x7B).
7.	Nama Jurnal	Computer Science, Computation and Language
	Tahun Penerbitan	2023
	Judul	Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models.
	Penulis	Lochan Basyal, Mihir Sanghvi [31]
	Metode Penelitian	Model yang digunakan akan diujicobakan untuk meringkas 25 dataset yang berisi berita dari CNN/Daily Mail dan Xsum menggunakan NVIDIA T4 dengan mengutilisasi LangChain dan Hugging Chain.
	Hasil	Dilakukan pembuatan model peringkasan teks dengan menggunakan MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models dengan hasil tertinggi didapatkan oleh model OpenAI Chat-GPT Model dengan hasil akurasi dataset CNN/Daily Mail pada ROUGE-1 sebesar 0.272, ROUGE-2 sebesar 0.096, dan ROUGE-L sebesar 0.255.

Pada tabel 2.1 menunjukan penelitian-penelitian terdahulu yang menjadi pondasi dari penelitian yang dilakukan. Beberapa penelitian sebelumnya menunjukan keunggulan model LLM Mistral 7B dibandingkan dengan model lainnya [23] [27] [28] [29] [30], terdapat penelitian yang menunjukan efektivitas dari penggunaan instruction tuning terhadap sebuah model LLM [26], dan terdapat penelitian yang menggunakan berbagai LLM untuk dibandingkan dalam peringkasan teks [31]. Terdapat penelitian yang memperkenalkan model Mistral 7B. penelitian ini menunjukan bahwa model Mistral 7B mampu menandingi model 13B pada setiap tolak ukur dan 34B pada penilaian matematika dan *code generation.* Peneliti juga membandingkan model Mistral 7B dengan model 7B lainnya dan didapatkan hasil model Mistral 7B mengalahkan model lainnya pada

11 dari 12 tes tolak ukur yang dilakukan. hasil juga menunjukan model Mistral 7B dapat menjawab pertanyaan dengan jawaban yang jelas dengan nilai evaluasi *precision* 99.4% dan *recall* 95.6% yang berarti menunjukan hasil yang positif, sedangkan model lainnya menolak untuk menjawab pertanyaan yang diberikan [23].

Penelitian menunjukan bahwa penggunaan *instruction tuning* dengan model yang dikembangkan bernama *InstructPTS* yang membantu dalam melakukan peringkasan judul produk. Pada penelitian ini, hasil model FLAN-T5-XL yang menggunakan *InstructPTS* mendapatkan nilai yang tertinggi pada setiap aspek penilaian, dibandingkan dengan model lainnya. Penelitian ini juga menunjukan hasil ringkasan judul produk menggunakan *InstructPTS* mencapai 96% telah sesuai dengan kriteria yang telah ditentukan [26].

. Pada penelitian terhadap *ontology engineering* menunjukan hasil Mistral 7B dapat melampaui hasil peringkasan yang dilakukan oleh GPT-4 dengan nilai *Recall* 48.75% dan F1-Score 0.8379 pada kelas, dan nilai *Recall* 45% dan F1-Score 0.4864. Diketahui bahwa hasil peringkasan Mistral 7B telah memberikan hasil yang lebih baik dibandingkan model GPT-4, tetapi perlu digarisbawahi diperlukan penyesuaian *fine tuning* yang sesuai untuk menyesuaikan pengetahuan yang spesifik terhadap hal tertentu [27]. Pada penelitian lainnya, Mistral 7B menunjukan performa yang lebih baik dibandingkan Llama 2-7B pada ujicoba pelengkapan riwayat klinis. Mistral 7B sebagai *base model opensource* juga mampu bersaing dengan *base model closesource* GPT-4 Turbo dengan nilai dengan nilai substansial (mean κ, 0.73 [95% CI: 0.67, 0.78] to 0.77 [95% CI: 0.71, 0.82]) [28].

Pada penelitian lainnya menunjukan model Mistral 7B yang telah dilatih dengan menggunakan dataset dengan ukuran 200k, 400k, dan 700k dengan menggunakan RTX 4090 selama 16 jam. Ketiga model tersebut diujicobakan pada OpenEval dan Mistral 7B dengan ukuran 200k mampu menyaingi 57 *base model* lainnya dan mendapatkan peringkat satu dengan hasil akhir evaluasi 0.660. Mistral 7B dengan ukuran 200k mendapatkan performa 35% lebih baik dibandingkan dengan juara 2 yang menggunakan Qwen-14B *based submission.* [29]. Pada penelitian pembuatan

model LLM untuk de-identifikasi data pasien, Mistral 7B menunjukan kinerja yang tinggi dengan nilai akhir F1-Score 0,9673. Penelitian menunjukan bahwa model Mistral 7B mendapatkan nilai evaluasi tertinggi dibandingkan Llama 2 dengan nilai akhir 0.8750 dan Mistral8x7B dengan nilai akhir 0.8686. Ujicoba di tingkat catatan klinis menunjukan Mistral 7B meningkat dengan hasil akhir 0.9915, sedangkan kedua model lainnya mengalami penurunan performa [30].

Terdapat penelitian lainnya yang menunjukan hasil dari peringkasan teks dengan menggunakan LLM MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT dengan menggunakan dataset CNN/Daily Mail. Penelitian ini menunjukan hasil dengan model terunggul didapatkan oleh model OpenAI Chat-gpt dengan hasil ROUGE-1 sebesar 0.272, ROUGE-2 sebesar 0.096, dan ROUGE-L sebesar 0.255 [31].

Berdasarkan dari penelitian-penelitian sebelumnya, model Mistral 7B konsisten menunjukan performa yang baik dibandingkan dengan model-model lainnya. Penggunaan *instruction tuning* dapat membantu dalam memberikan hasil yang sesuai dengan kebutuhan dan meningkatkan akurasi dari hasil akhir peringkasan. Oleh karena itu, Mistral 7B akan digunakan sebagai model untuk peringkasan teks dengan menggunakan *instruction tuning* untuk membantu dalam memberikan hasil yang sesuai dengan kebutuhan.

2.2 Tinjauan Teori

2.2.1 Teknologi Informasi

Teknologi informasi mengacu pada teknologi yang mencakup pengembangan, pemeliharaan, dan penggunaan sistem komputer, perangkat lunak, dan jaringan. Tujuan utamanya adalah untuk memproses dan mendistribusikan data secara efektif. perkembangan teknologi telah membawa tranformasi yang signifikan terhadap berbagai bidang, seperti Kesehatan, bisnism, Pendidikan, Kesehatan dan lainnya. Dengan teknologi informasi seperti pengolahan data besar (*big data*), *artificial intelliegnnce, cloud computing* merevolusikan cara manusia dalam mengelola informasi. [31].

2.2.2 Overload Information

Overload information atau kelebihan data merupakan fenomena dimana informasi yang ada melebihi kapasitas seseorang untuk memprosesnya. Dengan terlalu banyaknya informasi, dapat menyebabkan kebingungan dalam pencarian informasi yang sesuai dengan kebutuhan, dan kesulitan dalam pengambilan Keputusan. *Overload information* dapat terjadi karena teknologi informasi dan internet yang sangat berkembang sehingga meningkatkan jumlah informasi yang tersedia. Setiap tahun informasi selalu bertambah dan terus meningkat pertambahannya, mulai dari media sosial, informasi berita, hingga penelitian-penelitian yang dilakukan. Semakin banyaknya informasi memberikan dampak buruk dalam pencarian informasi yang spesifik. Informasi akan menumpuk satu dengan yang lainnya menyebabkan informasi akan menjadi terlalu banyak untuk dicerna menyebabkan fenomena kelebihan data di masyarakat. Pencarian informasi diantara informasi yang berlebihan menyebabkan kebingungan dalam pengambilan keputusan berdasarkan informasi yang dibaca [11].

2.2.3 Peringkasan Teks

Peringkasan teks merupakan proses untuk memadatkan teks yang panjang menjadi versi yang lebih pendek dengan tetap mempertahankan informasi penting dan makna keseluruhan. Peringkasan teks digunakan sebagai solusi dalam mengatasi kelebihan data yang terjadi di masyarakat. Seiring bertambahnya jumlah ada diperlukan metode untuk membaca data secara cepat dan efisien. Tujuan dari peringkasan teks adalah untuk mempercepat dalam pemahaman informasi, mengurangi waktu, dan mempermudah dalam pengambilan keputusan [15]. Terdapat dua tipe dalam peringkasan teks berdasarkan dari informasi yang disimpulkan.

2.2.3.1 Extractive Summarization

Extractive summarization adalah teknik dalam pemrosesan bahasa alami yang bertujuan untuk menghasilkan ringkasan dengan memilih kalimat-kalimat penting dari teks asli. Metode ini mengandalkan algoritma untuk menilai relevansi dan signifikansi kalimat, sehingga dapat menyusun informasi yang padat dan informatif tanpa mengubah makna aslinya [31].

2.2.3.1 Abstractive Summarization

Abstractive summatization merupakan metode pembuatan ringkasan ringkasan yang mungkin berisi kata, frasa, atau kalimat yang tidak yang tidak ada dalam teks sumber. Pendekatan ini bergantung pada pemahaman konteks dan menghasilkan bahasa yang mirip manusia untuk menyampaikan menyampaikan ide-ide utama [32].

2.2.4 LLM

LLM (*Large Language Model*) merupakan salah satu model *Artificial Intelligence* (AI) yang memanfaatkan teknik *deep learning* berbasis arsitektur *neural network* untuk memproses dan menghasilkan output berupa teks (*text generation*) berdasarkan input yang diberikan. Model ini dilatih menggunakan sejumlah besar data teks yang mencakup berbagai domain pengetahuan, sehingga mampu mempelajari pola linguistik, struktur tata bahasa, konteks semantik, serta relasi antar-entitas dalam bahasa alami. Dengan kapasitas pemrosesan yang tinggi, LLM dapat menghasilkan teks yang tidak hanya koheren dan relevan, tetapi juga adaptif terhadap berbagai gaya penulisan dan kebutuhan pengguna. LLM memiliki kemampuan dalam pengolahan bahasa, seperti melakukan pembuatan teks, menjawab pertanyaan, melakukan peringkasan teks, dan fungsi lainnya yang berhubungan dengan analisis dan bahasa [20] [33].

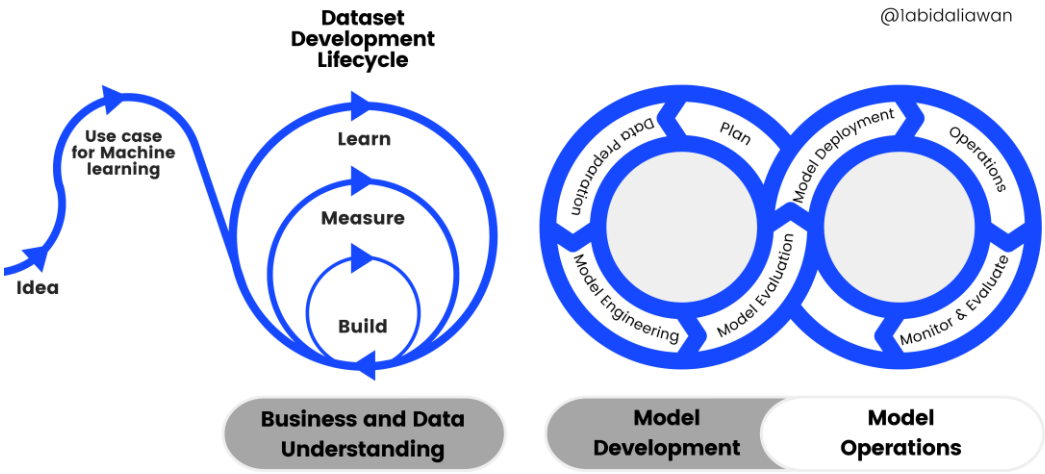
2.2.5 Instruction Tuning

Instruction tuning merupakan salah satu jenis *tuning* untuk LLM yang meningkatkan kemampuan model dalam mengikuti instruksi dan melaksanakan tugas tertentu dengan efektif. Pada penerapannya, Model akan diberikan instruksi kompleks sehingga meningkatkan pemahaman model terhadap data yang digunakan dan menyesuaikan *output* model berdasarkan intruksi yang diberikan. Dengan menggunakan *instruction tuning* kepada model dapat menyempurnakan dan meningkatkan kemampuan model yang telah dilatih sebelumnya, sehingga dapat menyesuaikan scenario yang kompleks dan beragam [34].

2.3 Framework yang Digunakan

2.3.1 CRISP-ML

CRISP-ML (*Cross-Industry Standard Process for Machine Learning)* merupakan kerangka kerja yang dibuat untuk memberikan panduan pengembangan penerapan *machine learning [35]*. CRISP-ML berasal dari model kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining)* yang disesuaikan untuk memenuhi kebutuhan interpretasi akan penerapan *machine learning* yang tidak dapat dilingkupi oleh kerangka kerja CRISP-DM. Kerangka kerja CRISP-ML memiliki kelebihan yaitu memiliki aspek keamanan jaminan kualitas, sehingga model dapat terus menerus dipastikan sesuai dengan kebutuhan bisnis dan memiliki performa yang sesuai [36]. Pada gambar 2.1 menunjukan kerangka kerja CRISP-ML memiliki fleksibilitas dalam setiap langkah proses, memungkinkan untuk kembali ke tahap sebelumnya jika terdapat perubahan dalam data atau jika ada kebutuhan baru yang muncul Terdapat 6 tahapan pada CRISP-ML, dengan setiap langkah bertujuan untuk memastikan hasil projek dapat sesuai dengan kebutuhan yang diharapkan [37].



Gambar 2. 1 CRISP-ML Life Cycle Process [38]

Terdapat 6 tahapan dalam CRISP-ML [36][37], antara lain:

1. **Business and Data Understanding**

Tahap pertama dalam CRISP-ML adalah memahami tujuan bisnis, kemudian mengumpulkan data yang diperlukan dan melakukan penilaian terhadap kelayakan proyek. Pada tahap ini bertujuan untuk memastikan penelitian menggunakan model *machine learning* dapat memberikan hasil yang diperlukan untuk menjawab permasalahan yang ada. Langkah ini mencakup mengidentifikasi keperluan bisnis, menganalisis kebutuhan penelitian, dan memastikan data yang digunakan selaras dengan tujuan penelitian.

2. **Data Preparation**

Tahap kedua dalam kerangka kerja CRISP-ML adalah pengolahan data. Pada fase ini, data yang diperlukan akan diproses untuk memastikan kualitas data sebelum memasuki tahap permodelan. Proses pengolahan ini melibatkan berbagai metode preprocessing, termasuk pembersihan data (*data cleansing*) dan pemilihan subset data yang relevan. Tujuan dari tahap ini adalah untuk memastikan bahwa data yang dipilih sesuai dengan tujuan penelitian serta untuk meningkatkan akurasi hasil dari model machine learning yang akan dikembangkan.

3. **Modelling**

Tahap ketiga adalah membuat model yang diperlukan untuk penelitian. Pada tahap ini, dilakukan pemilihan terhadap model LLM yang sesuai dengan kebutuhan penelitian. Model yang dipilih akan dilatih menggunakan data yang telah disiapkan. Proses ini mencakup eksperimen dengan berbagai model, fitur, dan parameter untuk mencapai kinerja optimal sesuai dengan kriteria yang telah ditentukan dalam penelitian.

4. **Evaluation**

Tahap selanjutnya setelah model dilatih adalah evaluasi, yang bertujuan untuk memastikan bahwa model tersebut memenuhi tujuan penelitian yang telah ditetapkan sebelumnya. Pada tahap ini, model yang telah dilatih akan diuji menggunakan data uji yang terpisah dari data pelatihan. Pengujian ini dilakukan untuk menilai kinerja model berdasarkan

parameter yang telah ditentukan sebelumnya. Evaluasi model melibatkan penggunaan berbagai metrik penilaian yang relevan, yang dapat memberikan gambaran yang jelas tentang seberapa baik model dalam menghasilkan prediksi yang akurat. Dengan adanya metrik penilaian ini, peneliti dapat menilai sejauh mana model mampu memenuhi kebutuhan bisnis dan tujuan penelitian yang telah ditetapkan. Hasil dari tahap evaluasi ini akan menjadi dasar untuk keputusan selanjutnya, termasuk apakah model siap untuk diterapkan dalam lingkungan produksi atau perlu dilakukan perbaikan lebih lanjut.

5. **Deployment**

   Setelah model dipastikan memenuhi kriteria dan sesuai dengan kebutuhan bisnis yang ditetapkan, model tersebut sudah dapat diterapkan. Tahap *deployment* akan mengintegrasikan model terhadap sistem yang ada dan mempersiapkan struktur untuk operasionalisasi model. Tahap ini dapat diimplementasikan dengan beberapa sistem yang ada.

6. **Monitoring**

   Tahapan terakhir adalah memastikan model yang diterapkan dipantau dan diawasi secara berkala untuk mendektesi perubahan data yang ada dapat mempengaruhi kinerja model. Tahap ini berfungsi untuk memastikan model agar terus memberikan hasil yang sesuai dengan kebutuhan yang telah ditetapkan dalam jangka waktu yang lama. Pemeliharaan dapat dilakukan dengan pembaruan model secara berkala untuk menyesuaikan kebutuhan bisnis.

2.4 Model LLM yang digunakan

2.4.1 Mistral 7B

Mistral 7B merupakan *large language model* dengan 7 miliar parameter yang dibuat oleh tim Mistral AI. Mistral 7B menggunakan teknik *deep learning* pada arstitektur *transformer* untuk memproses dan menghasilkan teks sehingga memiliki kinerja dan efisiensi yang tinggi dalam berbagai tugas *natural language processing.* Pada ujicoba yang

dilakukan, Mistral 7B lebih unggul dibandingkan dengan model Mistral 13B, dan dapat melampaui hasil Llama 34B sebagai model terbaik di 34B pada tes matematika dan *code generation*. Mistral 7B menggunakan perhitungan SWA (*Sliding Window Attention)* untuk membantu dalam menangani urutan yang lebih panjang secara lebih efektif dengan biaya komputasi yang lebih rendah, sehingga mengurangi batasan umum dalam LLM [23]. Dengan menggunakan SWA memungkinkan *Transformers* untuk menangkap konteks yang relevan dan, melalui beberapa lapisan, membangun pemahaman yang komprehensif dari keseluruhan urutan. Hal ini serupa dengan proses pembelajaran fitur hirarkis dalam Convolutional Neural Network (CNN) [39]. Model Mistral 7B merupakan model *open-source* yang dirilis dengan lisensi Apache 2.0 yang dapat digunakan secara bebas untuk keperluan kormesial maupun penelitian tanpa batasan ketat. Model ini juga dapat digunakan melalui *tools* Ollama atau dapat diinstal langsung melalui Wesbite Hugging Face.

2.5 Teori tentang Tools/Software yang digunakan

2.5.1 Python

Gambar 2. 2 Logo Bahasa Pemrograman Python [41]

Python adalah bahasa pemrograman tingkat tinggi yang dikenal karena keserbagunaannya dan pendekatan berorientasi objek *(object-oriented*), yang membuatnya mudah dipahami dan dipelajari. Salah satu keunggulan utama Python adalah sintaksisnya yang jelas dan mudah dimengerti, sehingga memungkinkan pengembang untuk menulis kode dengan lebih efisien dan mengurangi kemungkinan kesalahan. Python

dilengkapi dengan berbagai pustaka (*library*) yang menyediakan modul-modul lengkap untuk mendukung berbagai aplikasi, mulai dari pengembangan web hingga analisis data. Selain itu, Python dapat dijalankan di berbagai platform, termasuk Windows, macOS, dan Linux, menjadikannya pilihan yang fleksibel untuk berbagai jenis proyek. Python juga bersifat *open source*, yang berarti bahwa kode sumbernya dapat diakses dan dimodifikasi oleh pengguna [40] [41].

2.5.2 Jupyter Notebook

Jupyter Notebook adalah pemrograman interaktif berbasis web yang banyak digunakan dalam ilmu data dan pendidikan. Ini memungkinkan pengguna untuk menggabungkan kode, dokumentasi, dan visualisasi dalam satu antarmuka, meningkatkan pengalaman belajar dan mengajar. Jupyter Notebook memiliki berbagai fitur dapat menyimpan file dalam berbagai jenis file seperti pdf, ipynb, html, tex, md, asciidoc dan lainnya , dan memungkinkan untuk transfer kode model dari satu *instance* ke *instance* lainnya untuk melatih ulang model sehingga memudahkan pengguna dalam mengkonfigurasi dan mengatur alur kerja untuk berbagai keperluan, seperti analisis data, *machine learning,* dan komputasi jurnalisme Jupyter Notebook juga memiliki tampilan interaktif yang membantu dalam menjalankan kode secara langsung dan dapat langsung melihat hasilnya. Pada kode yang error dapat dilihat informasi mengenai error yang ada dan pada line berapa di kode tersebut yang menyebabkan error. Selain itu fitur *markdown* yang dapat memberikan sebuah catatan sehingga dapat dilakukan dokumentasi kode dengan baik, setiap kode dapat diberikan catatan mengenai fungsi maupun informasi tambahan yang diperlukan [42].

2.5.3 Ollama

Gambar 2. 3 Logo Kerangka Kerja Ollama [43]

Ollama adalah kerangka kerja sumber terbuka yang dirancang untuk pengelolaan dan eksekusi model LLM di mesin lokal. Platform ini menyediakan lingkungan yang ringan dan dapat diperluas, memungkinkan pengembang untuk membuat, menjalankan, dan mengelola model-model ini secara efisien. Ollama memfasilitasi berbagai tugas, seperti pemrosesan bahasa alami dan analisis data, sambil memastikan privasi data dengan menjaga operasi tetap lokal. Dengan menggabungkan bobot model, konfigurasi, dan data ke dalam satu paket yang didefinisikan oleh Modelfile, Ollama memudahkan pengguna dalam mengelola dan menjalankan model, mengurangi ketergantungan pada *cloud service*, serta tampilan *GUI* sehingga membantu dalam penelitian menggunakan LLM [43] [44].

2.6 Evaluasi yang digunakan

2.6.1ROUGE Score

Recall-Oriented for Gisting Evaluation (ROUGE) Score merupakan sebuah skor evaluasi untuk menilai tumpeng tindih antar *n-gram* (urutan kata) yang dihasilkan dengan ringkasan referensi. ROUGE Score sering digunakan dan menjadi standar metrik untuk evaluasi *generated output* yang dihasilkan model terhadap referensi yang digunakan. Terdapat evaluasi

recall yang biasa digunakan untuk menghitung kesamaan kata antara hasil *generated output* dan referensi yang diberikan. Hal ini menjadi kurang baik dikarenakan hasil dari model yang panjang tetap dapat membaca kesamaan kata yang ada pada referensi, sehingga walaupun banyak kata yang tidak sesuai akan tetap terhitung sebagai nilai *recall* yang baik. ROUGE Score menghitung nilai berdasarkan dari urutan kata antara hasil model dengan referensi untuk melihat kesamaan dalam dataset tersebut [20] [31].

Terdapat tiga jenis ROUGE Score yang digunakan, antara lain:

1. ROUGE 1

   ROUGE 1 merupakan metrik evaluasi yang mengukur nilai tumpeng tindih (*overlap)* dengan antara *unigram* (kata tunggal) yang dihasilkan melalui *generated output* dari model dengan ringkasan referensi. ROUGE 1 akan membandingkan setiap kata yang ada pada setiap kata tunggal yang dihasilkan dengan referensi tanpa memperdulikan urutan kata yang ada, sehingga dapat menilai akurasi kecocokan data antara kata yang dibuat oleh model dengan referensi secara keseluruhan.

   Berikut merupakan rumus perhitungan ROUGE 1:

$$ROUGE1_{Recall} = \frac{Unigram_{match}}{Unigram_{ref}}$$

$$ROUGE\ 1_{Precision} = \frac{Unigram_{match}}{Unigram_{gen}}$$

$$ROUGE1_{f1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

2. ROUGE 2

   ROUGE 2 merupakan metrik evaluasi yang menungkur nilai tumpah tindih (*overlap)* dengan antara nilai *bigram* (2 kata yang berurutan) yang dihasilkan *generated output* dari model dengan ringkasan referensi. Penilaian ini digunakan untuk menilai 2 urutan kata

berdasarkan dari *generated output* dari hasil model dengan referensi dengan tujuan untuk evaluasi terhadap informasi konstektual dan koherensi berdasarkan 2 urutan kata.

Adapun rumus ROUGE 2 yaitu:

$$ROUGE2_{Recall} = \frac{Bigram_{match}}{Bigram_{ref}}$$

$$ROUGE\ 2_{Precision} = \frac{Bigram_{match}}{Bigram_{gen}}$$

$$ROUGE2_{f1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (2)$$

3. ROUGE L

ROUGE L merupakan metrik evaluasi yang berorientasi pada LCS (*Long Common Subsequencce)* atau urutan terpanjang yang muncul dalam ringkasan hasil model dengan referensi. Dengan menggunakan ROUGE L dapat diketahui apakah sebuah subsekuens yang terdapat pada referensi yang digunakan dapat diringkas dengan mempertahankan informasi dengan makna yang sama. LCS tidak secara langsung mempertahankan urutan kata yang ada pada subsekuens sehingga informasi tidak harus disampaikan secara berurutan

Terdapat rumus ROUGE L yaitu:

$$ROUGEL_{Recall} = \frac{LCS(X,Y)}{|X|}$$

$$ROUGE\ L_{Precision} = \frac{LCS(X,Y)}{|Y|}$$

$$ROUGEL_{f1} = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{Recall + \beta^2 \cdot Precision} \quad (3)$$

v

Dengan penjelasan:

$X =$ Teks referensi

$Y =$ Teks hasil model

β = Faktor penyeimbang recall dan precision

2.6.2 Human Evaluation

Human Evaluation merupakan evaluasi manual yang dilakukan manusia dengan mengukur hasil *generated output* dari hasil model. Pada penilaian metrik otomatis seringkali tidak dapat sepenuhnya diukur sepenuhnya dan tidak selalu sejalan dengan harapan manusia. Pada evaluasi peringkasan teks yang telah dilakukan, sulit untuk mengukur nilai subjektivitas terhadap peringkasan yang telah dibuat. Penilaian seperti kecukupan makna, kesesuaian konteks, *naturalness* dan koherensi kata yang sulit dilakukan melalui metrik otomatis seperti BLEU Score mauoun ROUGE Score. *Human evaluation* menjadi evaluasi manual yang dilakukan dengan menargetkan populasi tertentu untuk diberikan pertanyaan dengan berbagai jenis skala penilaian. *Human evauation* memiliki panduan umum dalam pelakasanaannya, seperti target peserta evaluasi yang harus memiliki pemahaman terhadap konteks yang dievaluasikan. Untuk memastikan kualitas dari evaluasi, diperlukan pertanyaan uji dengan penilaian berdasarkan metrik penilaian yang dipilih [45]. Terdapat beberapa tahapan dalam pelaksanaan *human evaluation*, antara lain:

1. Perancangan Evaluasi

   Tahap perancangan evaluasi adalah penentuan dasar-dasar dari evaluasi yang ingin dinilai. Tahap ini akan dilakukan penentuan kriteria penilaian, seperti menggunakan *likert scale* yang menilai dari 1-5, *ranking* yang mengurutkan sebuah nilai komparasi tertentu, kemudian akan dipilih pertanyaan yang sesuai dengan kebutuhan evaluasi. Pertanyaan yang diajukan kepada evaluator harus memiliki kepentingan yang sesuai dengan tujuan penelitian.
2. Pemilihan Evaluator

Pemilihan evaluator sangat penting untuk evaluasi yang dilakukan. evaluator berfungsi untuk mengisi pertanyaan yang disediakan selama sesi evaluasi, maka sangat penting untuk memastikan bahwa pihak evaluator merupakan objek maupun pihak yang terlibat dampak dari penelitian yang dilakukan.

3. Analisis Statistik

   Tahap terakhir dari *human evaluation* adalah melakukan analisis statistik terhadap pertanyaan yang telah dijawab oleh evaluator

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA