# Halaman Judul

PENGEMBANGAN *LARGE LANGUAGE* *MODEL*

UNTUK MENJAWAB PERTANYAAN TERKAIT AKADEMIK

DI UNIVERSITAS SYIAH KUALA DENGAN

METODE *FINE-TUNING* DAN

*RETRIEVAL-AUGMENTED GENERATION*

TESIS

Diajukan untuk melengkapi tugas-tugas dan

memenuhi syarat-syarat guna pelaksanaan penelitian Tesis

**Oleh:**

HARY RACHMAT

2108207010009



PROGRAM STUDI MAGISTER KECERDASAN BUATAN

DEPARTEMEN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA, BANDA ACEH

AGUSTUS, 2024

# Halaman Pengesahan Laporan Tesis

PENGESAHAN TESIS

**PENGEMBANGAN *LARGE LANGUAGE MODEL* UNTUK MENJAWAB PERTANYAAN TERKAIT AKADEMIK DI UNIVERSITAS SYIAH KUALA DENGAN METODE *FINE-TUNING* DAN *RETRIEVAL-AUGMENTED GENERATION***

***DEVELOPMENT OF LARGE LANGUAGE MODEL TO ANSWER ACADEMIC-RELATED QUESTIONS AT SYIAH KUALA UNIVERSITY USING FINE-TUNING AND RETRIEVAL-AUGMENTED GENERATION METHODS***

Oleh:

Nama : Hary Rachmat

NPM : 2108207010009

Program Studi : Magister Kecerdasan Buatan

Menyetujui:

|  |  |  |
| --- | --- | --- |
| Pembimbing I, | Pembimbing II, | |
|  |  | |
| Prof. Dr. Ir. Hammam Riza, M.Sc. IPU | Prof. Dr. Taufik Fuadi Abidin, S.Si., M.Tech | |
| NIP. 196208081987111001 | NIP. 197010081994031002 | |
| Mengetahui: | | |
| Dekan Fakultas MIPA  Universitas Syiah Kuala | | Koordinator Program Studi  Magister Kecerdasan Buatan FMIPA  Universitas Syiah Kuala, |
|  | | |
| Prof. Dr. Taufik Fuadi Abidin, S.Si., M.Tech | | Dr. Nizamuddin, M.Info.Sc. |
| NIP. 197010081994031002 | | NIP. 197108241996031001 |

# Pernyataan Bebas Plagiasi

PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini,

Nama lengkap : Hary Rachmat

Tempat/tanggal lahir : Banda Aceh/18 April 1997

NPM : 2108207010009

Program Studi : Magister Kecerdasan Buatan

Jurusan : Informatika

Fakultas : MIPA

Judul Tugas Tesis\* : Pengembangan *Large Language Model* untuk Menjawab Pertanyaan terkait Akademik di Universitas Syiah Kuala dengan Metode *Fine-Tuning* dan *Retrieval-Augmented Genera*tion

Menyatakan dengan sesungguhnya bahwa Laporan Tesis saya dengan judul seperti di atas adalah **hasil karya saya sendiri** bersama dosen pembimbing dan **bebas plagiasi**.

Jika ternyata dikemudian hari terbukti bahwa Laporan Tesis merupakan hasil plagiasi, saya bersedia menerima sanksi yang berlaku di Universitas Syiah Kuala.

Banda Aceh, 31 Agustus 2024

Yang menyatakan,

Hary Rachmat

NPM. 2108207010009

**PERNYATAAN PERSETUJUAN PUBLIKASI DAN EMBARGO**

**KARYA TULIS CIVITAS AKADEMIKA USK**

Saya yang bertanda tangan di bawah ini,

Nama lengkap : Hary Rachmat

NPM : 2108207010009

Judul Tugas Tesis : Pengembangan *Large Language Model* untuk Menjawab Pertanyaan terkait Akademik di Universitas Syiah Kuala dengan Metode *Fine-Tuning* dan *Retrieval-Augmented Genera*tion

Fakultas : MIPA

Jurusan : Informatika

Program Studi : Magister Kecerdasan Buatan

menyetujui:

( ) untuk mengunggah softcopy Tugas Akhir/Tesis saya diatas secara penuh atau full text di Repository Perpustakaan USK dan diakses atau terbaca lewat mesin pencari internet secara publik

( ) untuk mengunggah softcopy Tugas Akhir/Tesis kami diatas secara parsial atau hanya bagian Tertentu (misalnya Cover, Lembar Pengesahan, Abstrak, Daftar Isi, Pendahuluan dan Kesimpulan) ke Repository Perpustakaan USK dan diakses atau terbaca lewat mesin pencari internet secara publik

(✓) untuk **tidak membolehkan** sama sekali bagian Tugas Akhir/Tesis diakses online sampai dengan **selamanya/terhingga** denganalasan :

**Akan dipublikasi di Jurnal Internasional.**

dan setuju memberikan atau mengunggah softcopy Tugas Akhir/Tesis kepada Perpustakaan USK secara penuh dalam bentuk PDF untuk disimpan secara private ke dalam Repositori yang tidak bisa dijangkau oleh mesin pencari (mis. Google).

Demikian pernyataan ini dibuat untuk dipergunakan sebagaimana mestinya.

Banda Aceh, 31 Agustus 2024

Yang menyatakan,

Hary Rachmat

NPM. 2108207010009

|  |  |
| --- | --- |
| Mengetahui |  |
| Pembimbing | Koordinator Program Studi |
|  |  |
| Prof. Dr. Ir. Hammam Riza, M.Sc. IPU | Dr. Nizamuddin, M.Info.Sc. |
| NIP. 196208081987111001 | NIP. 197108241996031001 |

ABSTRAK

Selama ini, informasi terkait akademik di Universitas Syiah Kuala (USK), selain di informasikan dalam bentuk website juga dirangkum dalam bentuk Frequently Asked Question (FAQ). Informasi dalam bentuk website dan FAQ tidak interaktif sehingga informasi tertentu harus dicari dalam berita web atau daftar FAQ tersebut. Oleh karena itu, dibutuhkan alternatif cara penyampaian informasi yang lebih interaktif menggunakan chatbot. Chatbot dapat dibangun dengan menggunakan Large Language Model (LLM) seperti Mistral 7B. Mistral 7B adalah model bahasa besar yang dapat diterapkan untuk menjawab pertanyaan seperti informasi akademik menggunakan data yang dikumpulkan dari universitas. Metode fine-tuning dengan Teknik QLoRA dan RAG dapat digunakan untuk melatih model dan mengambil informasi yang relevan dari sumber dokumen eksternal. Hasilnya kemudian dievaluasi menggunakan skor ROUGE. Jawaban dari model USK Mistral 7B memberikan hasil dengan skor >0,5 dengan 15 dari 56 pertanyaan menggunakan metode RAG dan metode fine-tuning diuji pada 20 pertanyaan menghasilkan nilai dengan skor >0,5. Pengujian juga dilakukan dengan pertanyaan yang berbeda tetapi memiliki makna yang sama dan didapatkan hasil tanggapan dengan skor ROUGE 0,4-0,5 dari pertanyaan yang diajukan. Dengan adanya model USK Mistral 7B ditemukan suatu cara yang dapat digunakan untuk membantu pengguna dalam memberikan informasi terkait akademik di USK secara interaktif.

Kata kunci: Large Language Model, Fine-tuning, RAG

ABSTRACT

Right now, academic information at Universitas Syiah Kuala (USK) is distributed on a website or summarized in the form of Frequently Asked Questions (FAQ). Information in the form of a website and FAQ is not interactive. Certain information must be searched from the web or FAQ list. Therefore, a more interactive way to get information using a chatbot is needed. Chatbots can be built using a Large Language Model (LLM) such as Mistral 7B. Mistral 7B is a large language model that can be applied to answer questions such as academic information using data collected from universities. The fine-tuning method with the QLoRA and RAG techniques can be used to train the model and retrieve relevant information from external document sources. The results are then evaluated using the ROUGE score. The answers from the USK Mistral 7B model gave results with a score of >0.5 on 15 out of 56 questions using the RAG method, and the fine-tuning method was tested on 20 questions, producing a value with a score of >0.5. Testing was also conducted with different questions that had the same meaning, and response results were obtained with a ROUGE score of 0.4-0.5 from the questions asked. Using the USK Mistral 7B model in a chatbot, academic information at USK can be shared interactively.

Keywords: Large Language Model, Fine-tuning, RAG

# Kata Pengantar

KATA PENGANTAR

Syukur Alhamdulillah dipanjatkan ke hadirat Allah SWT yang telah melimpahkan rahmat-Nya sehingga Tesis yang berjudul “**Pengembangan *Large Language Model* untuk Menjawab Pertanyaan Terkait Akademik di Universitas Syiah Kuala dengan Metode *Fine-Tuning* dan *Retrieval-Augmented Genera*tion”** dapat diselesaikan.Selawat dan salam disanjungkan kepada Nabi Besar Muhammad SAW.

Tesis ini merupakan salah satu syarat yang harus dipenuhi untuk memperoleh gelar Magister di Departemen Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala. Penyelesaian penulisan Tesis ini tidak terlepas dari bantuan dan dorongan dari berbagai pihak, baik secara moril maupun materil. Pada kesempatan ini, ucapan terimakasih diucapkan kepada:

1. Bapak Dr. Nizamuddin, M.Info.Sc. selaku Koordinator Program Studi Magister Kecerdasan Buatan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala.
2. Bapak Prof. Dr. Ir. Hammam Riza, M.Sc. IPU, selaku Pembimbing I dan Bapak Prof. Dr. Taufik Fuadi Abidin, S.Si., M.Tech, selaku pembimbing II yang telah membimbing dan memberikan arahan dan bimbingan kepada penulis dalam penyelesaian Tesis ini.
3. Ayahanda dan Ibunda yang telah mendukung penulis dalam penyelesaian studi di Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala.
4. Semua pihak yang telah memberikan kontribusi baik moral maupun material sehingga penulis dapat menyelesaikan Laporan Tesis ini.

Semoga tulisan ini dapat bermanfaat bagi perkembangan ilmu pengetahuan.

Banda Aceh, 31 Agustus 2024

**Hary Rachmat**

NPM. 2108207010009

# Daftar Isi

DAFTAR ISI

*Halaman*

[Halaman Judul i](#_Toc178793507)

[Halaman Pengesahan Laporan Tesis ii](#_Toc178793508)

[Pernyataan Bebas Plagiasi ii](#_Toc178793509)

[Kata Pengantar vii](#_Toc178793510)

[Daftar Isi viii](#_Toc178793511)

[Daftar Tabel x](#_Toc178793512)

[Daftar Gambar xi](#_Toc178793513)

[Daftar Lampiran xii](#_Toc178793514)

[BAB I PENDAHULUAN](#_Toc178793515)

[1.1 Latar Belakang 1](#_Toc178793516)

[1.2 Rumusan Masalah 2](#_Toc178793517)

[1.3 Tujuan Penelitian 3](#_Toc178793518)

[1.4 Manfaat Penelitian 3](#_Toc178793519)

[BAB II TINJAUAN KEPUSTAKAAN](#_Toc178793520)

[2.1 Kecerdasan Buatan 4](#_Toc178793521)

[2.2 *Natural Language Processing* 4](#_Toc178793522)

[2.3 *Large Language Model* 6](#_Toc178793523)

[2.4 *Chatbot* 9](#_Toc178793524)

[2.5 *Model Mistral 7B* 9](#_Toc178793525)

[2.6 *Model Llama 2* 11](#_Toc178793526)

[2.7 *Fine-tuning* 11](#_Toc178793527)

[2.8 *Retrieval-Augmented Generation* 13](#_Toc178793528)

[2.9 Kuantisasi 14](#_Toc178793529)

[2.10 GPTQ: *Post-Training Quantization For GPT* 17](#_Toc178793530)

[2.11 GGUF: *GPT-Generated Unified Format* 18](#_Toc178793531)

[2.12 *Dataset* 18](#_Toc178793532)

[2.13 USK Dataset 20](#_Toc178793533)

[2.14 *Python* 20](#_Toc178793534)

[BAB III METODOLOGI PENELITIAN](#_Toc178793535)

[3.1 Tempat dan Waktu Penelitian 22](#_Toc178793536)

[3.2 Alat dan Bahan 22](#_Toc178793537)

[3.3 Metode Penelitian 22](#_Toc178793538)

[BAB IV HASIL DAN PEMBAHASAN](#_Toc178793539)

[4.1 Tampilan Chatbot dan Hasil Inferensi LLM 30](#_Toc178793540)

[4.2 Kondisi Dataset 31](#_Toc178793541)

[4.3 Kuantisasi dan Pelatihan Model 32](#_Toc178793542)

[4.4 Pendekatan LLM dengan metode RAG 35](#_Toc178793543)

[4.5 Hasil Pengujian dan Evaluasi Hasil Inferensi 36](#_Toc178793544)

[4.6 Kategori skor ROUGE 42](#_Toc178793545)

[4.7 Menghitung Evaluasi Sumber Daya 43](#_Toc178793546)

[4.8 Analisis Hasil 44](#_Toc178793547)

[BAB V KESIMPULAN DAN SARAN](#_Toc178793548)

[5.1 Kesimpulan 46](#_Toc178793549)

[5.2 Saran 46](#_Toc178793550)

[DAFTAR KEPUSTAKAAN 47](#_Toc178793551)

# Daftar Tabel

DAFTAR TABEL

*Halaman*

[Tabel 2.1. Kuantisasi 4-bit menggunakan llama.cpp (Sumber: Neves, 2023) 16](#_Toc178793468)

[Tabel 3.1. Alat dan Bahan 22](#_Toc178793469)

[Tabel 4.1. Konfigurasi *FastLanguageModel.from\_pretrained* 33](#_Toc178793470)

[Tabel 4.2. Konfigurasi FastLanguageModel.get\_peft\_model 34](#_Toc178793471)

[Tabel 4.3. Konfigurasi *SFTTrainer* 34](#_Toc178793472)

[Tabel 4.4. Konfigurasi *TrainingArguments* 35](#_Toc178793473)

[Tabel 4.5. Hasil dengan metode *fine-tuning* 37](#_Toc178793474)

[Tabel 4.6. Nilai skor ROUGE dengan Metode *Fine-tuning* 38](#_Toc178793475)

[Tabel 4.7. Hasil jawaban dengan Paraphrase pada pertanyaan 39](#_Toc178793476)

[Tabel 4.8. Hasil dengan metode RAG 41](#_Toc178793477)

[Tabel 4.9. Nilai skor ROUGE dengan Metode RAG 42](#_Toc178793478)

[Tabel 4.10. Tabel Kategori Nilai Metrik Rouge (Walker II, 2024) 43](#_Toc178793479)

[Tabel 4.11. Hitungan waktu pada model saat *fine-tuning* dan menjalankan RAG 43](#_Toc178793480)

# Daftar Gambar

DAFTAR GAMBAR

*Halaman*

[Gambar 2.1. Perbandingan performa pada Model Mistral 7B dengan Llama2 (Sumber: Jiang et al., 2023) 10](#_Toc178793407)

[Gambar 2.2. Alur *Fine-tuning* model (Sumber: Fatyanosa, 2020) 12](#_Toc178793408)

[Gambar 2.3. Arsitektur pada *Retrieval-augmented generation* (RAG) 14](#_Toc178793409)

[Gambar 2.4. Tipe data LLM terkuantisasi (Sumber: Neves, 2023) 15](#_Toc178793410)

[Gambar 2.5. Format numerik *bfloat16* (Sumber: Neves, 2023) 16](#_Toc178793411)

[Gambar 3.1. *Flowchart* Alur Penelitian 23](#_Toc178793412)

[Gambar 3.2. Alur Pelatihan pada LLM (Sumber: Benveniste, 2023) 24](#_Toc178793413)

[Gambar 3.3. Alur Pada RAG (Sumber: Benveniste, 2023) 26](#_Toc178793414)

[Gambar 3.4. Pipa pengideksan pada RAG (Sumber: Benveniste, 2023) 26](#_Toc178793415)

[Gambar 3.5. Desain UI Chatbot menggunakan Gradio (www.gradio.app, 2024) 29](#_Toc178793416)

[Gambar 4.1. Contoh Hasil inferensi Chatbot menggunakan desain Gradio UI 31](#_Toc178793417)

[Gambar 4.2. Proses *Fine-tuning* pada Model Mistral 7B 32](#_Toc178793418)

[Gambar 4.3. Pipeline RAG, selama fase kueri, mengambil konteks yang paling relevan dari perintah pengguna, meneruskannya ke *Large* *Language Model* 36](#_Toc178793419)

# Daftar Lampiran

DAFTAR LAMPIRAN

*Halaman*

[Lampiran 1. Sampel dataset Q&A di USK 50](#_Toc178793383)

[Lampiran 2. Sampel dataset untuk metode RAG 54](#_Toc178793384)

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# BAB I PENDAHULUAN

## Latar Belakang

Dengan kemajuan pesat kecerdasan buatan (AI), model bahasa besar (LLM) telah merevolusi pemrosesan bahasa alami (NLP) sehingga memungkinkan komputer untuk berinteraksi dengan teks dan bahasa. Salah satu contoh LLM adalah *Generative Pretrained Transformer* 4 (GPT-4) yang telah dilatih secara ekstensif pada data teks dalam jumlah besar sehingga memungkinkan GPT-4 melakukan analisis teks di berbagai domain. LLM dapat diaplikasikan pada berbagai bidang seperti bidang kesehatan, pendidikan, hukum, dan bidang-bidang lainnya dengan melakukan pra-pelatihan menggunakan data dalam bidang tersebut dalam jumlah besar.

Salah satu aplikasi chatbot yang populer saat ini adalah ChatGPT yang dikembangkan oleh OpenAI dan dirilis pada akhir tahun 2022 (Mohamadi et al., 2023). Berbagai penelitian berupaya mengeksplorasi kemampuan ChatGPT seperti dalam bidang medis yang dikembangkan untuk membantu dokumentasi klinis (Baker et al., 2024), mengklasifikasikan teks dalam bidang perbankan (Loukas et al., 2023) dan dalam bidang hukum untuk menentukan undang-undang mana yang berpotensi dilanggar (Trozze et al., 2023).

Selain LLM dengan sumber tertutup seperti GPT-4, beberapa penelitian menggunakan LLM sumber terbuka seperti yang dilakukan oleh Huang et al. (2023) dengan mengadaptasi LLM Llama ke domain hukum untuk membantu pengacara dalam membuat laporan teknis. Bhatti et al. (2023) melakukan *fine-tuning* pada LLM Llama 70B yang diberi nama “SM70” untuk menangani berbagai pertanyaan medis serta pengambilan keputusan klinis yang kompleks. Zhao et al. (2023) melakukan *fine-tuning* pada LLM Llama 7B yang diberi nama “Ophtha-LLaMA2” untuk membantu mendiagnosis penyakit mata yang akan memberikan dukungan keputusan bagi dokter. Barandoni et al. (2024) melakukan analisis komparatif LLM untuk mengekstraksi kebutuhan pelanggan travel dari postingan *TripAdvisor* dengan Memanfaatkan beragam model, termasuk model sumber terbuka seperti Mistral 7B dan sumber tertutup seperti GPT-4 dan Gemini. Hasil penelitan menyoroti kemanjuran LLM sumber terbuka, khususnya Mistral 7B, dalam mencapai kinerja yang sebanding dengan model sumber tertutup. Penelitian mengenai chatbot sebagai asisten virtual telah dilakukan oleh Jonatan & Igor (2023) untuk meningkatkan efisiensi layanan kepada pelanggan. Penelitian yang telah dipaparkan tersebut telah menunjukan potensi LLM untuk membantu pekerjaan dalam berbagai bidang.

Berdasarkan data yang di dapat dari website (data.usk.ac.id, 2024) jumlah mahasiswa yang mendaftar ke Universitas Syiah Kuala (USK) mengalami peningkatan setiap tahunnya. Untuk memberikan berbagai informasi seperti detail pendaftaran, uang kuliah, dan lain-lain USK menyediakan website dengan alamat usk.ac.id. sehingga mahasiswa dapat mencari informasi mengenai hal yang berkaitan dengan universitas. Pada saat ada informasi yang belum tersedia di website calon mahasiswa dapat bertanya secara langsung melalui sosial media seperti *direct message* (DM) melalui aplikasi *Instagram*, dan juga menjumpai staf perguruan tinggi secara langsung di bagian Unit Layanan Terpadu (ULT) atau bagian Hubungan Masyarakat (Humas). Penelitian ini akan mengkaji penerapan LLM sebagai asisten virtual dalam bentuk chatbot untuk memberikan informasi interaktif terkait administrasi akademik di USK.

## Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas, ada beberapa rumusan masalah yang perlu diteliti untuk menemukan solusinya. Adapun rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Pertanyaan dan jawaban terkait akademik selama ini disusun dalam bentuk *Frequently Asked Question* (FAQ). FAQ ini tidak interaktif sehingga jawaban atau informasi harus dicari dalam daftar FAQ tersebut. Oleh karena itu, dibutuhkan alternatif cara penyampaian informasi yang interaktif menggunakan chatbot.
2. Belum diketahui kinerja metode *fine-tuning* dan RAG dalam pengembangan chatbot untuk informasi akademik di USK.
3. Belum tersedia dokumen eksternal dalam bentuk pernyataan yang diperlukan untuk menggunakan RAG dalam pengembangan LLM di bidang akademik.

## Tujuan Penelitian

Adapun Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan LLM yang dapat memberikan informasi akademik di USK.
2. Mendapatkan metode dengan kinerja terbaik dalam pengembangan chatbot.
3. Mengumpulkan dokumen eksternal berupa informasi akademik di USK yang akan digunakan pada metode RAG dalam pengembangan LLM terkait akademik di USK.

## Manfaat Penelitian

Beberapa manfaat penelitian yang diharapkan dari penelitianini adalah sebagai berikut:

1. Informasi akademik di USK yang dapat diakses secara interaktif dengan memanfaatkan LLM.
2. Didapatkan metode terbaik dalam penerapan LLM di bidang Akademik.
3. Tersedia dokumen eksternal yang dapat digunakan dalam pengembangan LLM dengan metode RAG.
4. Peningkatan layanan serta kepuasan yang diberikan kepada mahasiswa di USK.
5. Menjadi kajian awal dengan pemanfaatan LLM serta mendapatkan ilmu pengetahuan dan informasi baru terkait variasi domain dalam penerapan LLM dalam bidang administrasi akademik di USK.

# BAB II TINJAUAN KEPUSTAKAAN

## Kecerdasan Buatan

Menurut Russel, Stuart J; Norvig (2010) sebuah perangkat pada komputer yang dapat memahami lingkungan di sekitarnya ialah kecerdasan buatan, tujuan dari tindakannya yaitu untuk memberikan suatu respons yang sesuai. Pengertian yang hampir sama juga diberikan oleh Minsky (1989) Menurutnya, kecerdasan buatan pada komputer merupakan suatu ilmu yang mempelajari cara membuat suatu program yang dapat melakukan suatu tindakan yang sama seperti yang manusia dapat lakukan.

Dengan dua pengertian tersebut, maka dapat ditarik kesimpulan bahwa Kecerdasan Buatan (*Artificial Intelligence*) merupakan suatu bidang ilmu yang digunakan dalam hal rancangan, pembangunan dan konstruksi pada satu mesin atau komputer pada program komputer hingga dapat dibentuk suatu kecerdasan seperti halnya manusia. Dalam hal ini kecerdasan berarti kemampuan dalam mengambil suatu tindakan, atau penyelesaian suatu masalah seperti halnya manusia yang dapat menggunakan kecerdasannya.

Kecerdasan Buatan meliputi suatu lingkup penelitian dari berbagai aspek seperti layaknya kecerdasan manusia dalam hal tingkat pengetahuan, pembelajaran, penalaran, perencanaan, pemrosesan bahasa, hingga kemampuan untuk memanipulasi suatu objek. Kecerdasan Buatan diharapkan dapat memiliki kecerdasan umum layaknya manusia.

## *Natural Language Processing*

*Natural language processing* (NLP) merupakan salah satu cabang dari kecerdasan buatan yang dapat membantu komputer memahami bahasa manusia baik secara lisan maupun tertulis, atau yang disebut dengan *natural language*.

Manfaat NLP mencakup:

1. Mengekstraksi tulisan, dimana NLP dapat merangkum sebuah tulisan dan menemukan inti dari tulisan tersebut.
2. Mendeteksi bahasa pada suatu tulisan lalu menerjemahkannya ke bahasa lain.
3. Menghasilkan konten secara otomatis berdasarkan teks yang tidak terstruktur.
4. Mendeteksi plagiarisme konten.
5. Mengidentifikasi suasana hati atau sentimen dari sebuah tulisan atau suara.
6. Mengubah perintah suara menjadi teks tertulis atau sebaliknya.

### Jenis-jenis NLP

Dikutip dari OpenGenus, ada 5 model NLP yang paling terbaru dan umum digunakan, yaitu:

1. BERT

BERT (*Bidirectional Encoder Representations from Transformers*) dikembangkan oleh Google pada tahun 2018. BERT dirancang untuk membantu komputer memahami makna bahasa yang ambigu di dalam suatu teks dengan cara memanfaatkan kata yang ada di dalamnya untuk membangun konteks. BERT dapat membaca teks dari dua arah sekaligus dan menganalisis bahasa tulisan menggunakan penyimpanan data pelatihan berlabel khusus. Data pelatihan berasal dari RNN pra-terlatih yang menggunakan sekumpulan teks tanpa label. Jadi, mesin menggunakan algoritma *machine learning* untuk menganalisis dan mengelompokkan data tanpa label (*unsupervised learning*), dan memberi label pada kalimat atau paragraf sebagai output.

1. GPT-3

GPT-3 (*Generative Pre-Trainer* generasi 3) merupakan model terbaru dari GPT dan GPT-2. GPT dibuat oleh OpenAI dan merupakan *machine learning* yang dilatih menggunakan data di dalam internet untuk menghasilkan segala jenis teks. GPT-3 hanya membutuhkan sedikit input untuk menghasilkan sebuah tulisan yang lengkap.

1. XLNet

XLNet bekerja menggunakan metode *autoregresif* (AR), yaitu menggunakan konteks suatu kata untuk memprediksi kata yang akan digunakan setelah atau sebelumnya. Untuk menghasilkan prediksi tersebut, XLNet berfokus pada permutasi atau disebut *Permutation Language Modelling* (PLM).

1. ALBERT

ALBERT adalah singkatan dari *A Lite BERT*. ALBERT memiliki desain yang mirip dengan BERT namun parameter yang digunakan jauh lebih sedikit. Misalnya, ALBERT menggunakan SOP (*Sentence Order Prediction*) dan mampu menangkap makna berbeda dari dua kata yang sama, misalnya buah ‘apel’ dan ‘apel’ pagi. Karena hal ini, ALBERT mampu memahami bahasa lebih akurat dibandingkan BERT.

1. RoBERTa

RoBERTa dikembangkan dari BERT dengan beberapa modifikasi, seperti menghapus fitur prediksi kalimat setelahnya (*Next Sentence Prediction*) dan dilatih dengan kumpulan kalimat yang lebih panjang (Revou.co, 2024).

## *Large Language Model*

Model bahasa besar menggunakan kecerdasan buatan untuk melakukan tugas pemrosesan bahasa alami (NLP) seperti terjemahan, pembuatan teks, menjawab pertanyaan, dan analisis sentimen. LLM menggunakan pembelajaran mendalam (*deep learning*) jenis pembelajaran mesin berbasis AI yang kompleks untuk memproses data dengan cara yang sangat mirip dengan interaksi alami manusia. LLM memungkinkan komputer untuk memahami, memanipulasi, dan menghasilkan teks dengan bahasa manusia. Seiring dengan semakin majunya teknologi ini, banyak tim mulai mengizinkan LLM untuk menyederhanakan dan mengotomatisasi proses yang berhubungan dengan manusia. LLM tidak hanya dapat menangani tugas rutin yang berulang seperti penjadwalan janji untuk bertemu atau FAQ, namun mereka juga dapat menangani tugas seperti pengambilan keputusan yang logis dan mengarahkan penelepon ke perwakilan yang tepat. Hal ini karena LLM memiliki akses ke data dalam jumlah besar sehingga berguna untuk tugas rutin dan lanjutan (Martin, 2023).

### Cara Kerja LLM

LLM dibangun menggunakan apa yang disebut arsitektur berbasis transformator. Transformator adalah jenis arsitektur “jaringan saraf” (seperti otak) yang dirancang oleh Google. Jaringan ini terdiri dari beberapa lapisan, yang semuanya bekerja sama untuk memecah teks menjadi bagian-bagian yang lebih kecil seperti kata atau karakter, yang disebut token, untuk menentukan hubungan dan makna antara setiap token.

Jaringan mencoba mengidentifikasi pola dalam data yang berurutan seperti kata-kata yang muncul setelah satu sama lain dalam sebuah kalimat untuk mengekstrak konteks dan makna dari pola tersebut. Satu lapisan dalam jaringan saraf LLM terdiri dari mekanisme perhatian. Hal ini memungkinkan model untuk fokus pada bagian teks tertentu untuk memahami konteks dan sentimennya. Hal ini membantu model merumuskan respons yang mirip manusia berdasarkan masukkan dari pengguna. Dapat dibayangkan arsitektur transformator ini seperti otak bayi yang baru lahir kemudian menerima stimulus dan secara alami mulai mengidentifikasi pola dan membentuk koneksi, oleh karena itu disebut “jaringan saraf” (Martin, 2023).

### Proses Pelatihan LLM

LLM menjalani proses “pelatihan” yang intensif dimana manusia memberi mereka data dalam jumlah besar untuk mengajari mereka cara memahami dan menghasilkan bahasa. Sama seperti bayi yang baru lahir mendengarkan suara orang dewasa dan mencoba memahami semuanya, LLM melakukannya dengan cara yang jauh lebih kompleks. LLM dilatih berdasarkan kumpulan data teks yang sangat besar, seperti buku, artikel, dan bahkan percakapan, namun kemudian harus disesuaikan untuk memberi hasil yang dapat disesuaikan tergantung pada tugas yang diberikan. Proses pelatihan LLM melewati dua fase utama: pra-pelatihan (*pretraining*) dan penyesuaian (*fine-tuning*) (Martin, 2023).

### Pra-pelatihan (*pretraining*)

Fase pertama yaitu pra-pelatihan LLM pada kumpulan data yang besar dan beragam. Jenis data ini dapat diambil dari internet dan dapat mencakup informasi dari media sosial, buku, artikel, dan situs web. Menggunakan data tersebut model mulai memahami hubungan antara berbagai token (Martin, 2023).

### Penyempurnaan (*fine-tuning*)

Fase berikutnya yaitu pembelajaran yang diawasi dimana manusia memberikan model berupa kumpulan data tambahan yang memungkinkannya bekerja lebih baik berdasarkan tugas tertentu. Misalnya, LLM dapat dilatih untuk digunakan dalam penawaran produk, petunjuk dan jawaban terkait merek tertentu. Hal ini membuat LLM dapat dengan mudah beradaptasi dengan tugas NLP yang berbeda. Proses pelatihan dua langkah ini memastikan LLM untuk dapat memahami masukan yang berbeda dan menghasilkan informasi yang akurat untuk pengguna (Martin, 2023).

### Pemelajaran ulang untuk kasus penggunaan spesifik

Salah satu manfaat utama dari model bahasa besar adalah dalam hal fleksibilitas dan adaptabilitasnya. Model yang sudah dilatih dengan korpus teks yang besar dapat dilatih ulang untuk kasus penggunaan spesifik, seperti pertanyaan dan jawaban serta dapat digunakan untuk menganalisis sentiment. Pemelajaran ulang merupakan proses pelatihan model dengan menggunakan dataset yang lebih kecil dan spesifik yang dapat digunakan untuk mengoptimalkan kinerjanya pada tugas tertentu. Proses ini membutuhkan waktu, namun memungkinkan suatu organisasi untuk dapat mengembangkan aplikasi pemrosesan bahasa yang akurat dan efektif dalam berbagai kasus. Ada banyak alat dan kerangka kerja yang memungkinkan pemelajaran ulang atau penyesuaian model bahasa besar. Banyak penyedia layanan kecerdasan buatan dapat menyesuaikan model bahasa besar yang sudah ada hanya dengan beberapa baris kode dan data pelatihan. Hal ini memungkinkan dalam menciptakan model bahasa yang dapat dikhususkan pada domain atau wilayah yang melebihi performa model bahasa generik atau umum pada kasus tertentu. Penyesuaian ini dapat dipersonalisasi untuk interaksi dengan pengguna serta meningkatkan pengalaman (www.wiz.ai, 2023).

### Manfaat pemelajaran ulang pada model bahasa besar

Manfaat pemelajaran ulang memungkinkan suatu perusahaan dapat membangun chatbot dalam mendukung pelayanan pada pelanggan. Dengan adanya chatbot, perusahaan dapat memberikan layanan kepada pelanggan mengenai produk dan pelayanan dengan cepat dan akurat. Chatbot juga dapat diberi sentuhan personalisasi sehingga dapat mencerminkan gaya dan nada merek perusahaan. Untuk mencapai hal ini, dapat dimulai dengan model bahasa besar yang sudah dilatih sebelumnya, sebagai contoh seperti model GPT-4 yang telah dilatih dengan jumlah data yang besar dan memiliki pemahaman yang kuat terhadap bahasa. Namun, model GPT-4 belum dilatih dengan data yang spesifik untuk industri atau produk tertentu, sehingga model belum memberikan performa yang sesuai dengan yang diharapkan untuk kasus penggunaan ini. Disinilah *fine-tuning* berperan penting, dengan melakukan *fine-tuning* model yang telah dilatih sebelumnya dapat dilatih ulang menggunakan data yang lebih spesifik pada suatu domain, seperti deskripsi produk, log dukungan pelanggan, dan informasi relevan lainnya. Dalam beberapa kasus, penyesuaian dapat dilakukan dengan menambahkan data baru ke dalam model pelatihan, yang lebih sederhana dan tidak membosankan daripada *fine-tuning*. Apapun metode yang digunakan, hasil akhirnya yaitu model dapat belajar mengenai informasi pada suatu domain seperti industri dan produk, serta dapat meningkatkan kinerja pada kasus penggunaan yang lebih spesifik. Dengan memanfaatkan metode *fine-tuning* pengguna dapat memberikan lebih banyak personalisasi dengan menggabungkan suara dan nada pada merek dalam domain tertentu dan menyediakan contoh mengenai cara perusahaan berkomunikasi dengan pelanggan. Hal ini membantu chatbot sehingga terdengar seperti merek yang memberikan pengalaman untuk pelanggan (www.wiz.ai, 2023).

## *Chatbot*

*Chatbot* adalah sebuah program berbasis kecerdasan buatan, yang dapat mensimulasikan percakapan atau obrolan dengan pengguna lain layaknya chat sesama pengguna melalui aplikasi pesan, situs web, aplikasi seluler, atau melalui telepon. teknologi ini juga dikenal sebagai asisten digital yang dapat memahami serta memproses permintaan pengguna, dan memberikan jawaban yang relevan dengan cepat. *Chatbot* digambarkan sebagai salah satu ekspresi dengan interaksi paling canggih dan menjanjikan antara manusia dan mesin. Di balik layar, cara kerja *chatbot* terbagi menjadi dua tugas inti yaitu analisis permintaan pengguna dan memberikan tanggapan. *Chatbot* memberikan tanggapan berdasarkan masukkan dari pengguna. Cara kerja *chatbot* terlihat sederhana dalam prakteknya namun sebenarnya cukup kompleks (Andryanto, 2021).

## *Model Mistral 7B*

Model *chatbot* Mistral 7B adalah model bahasa besar dengan parameter 7.3B yang dapat digunakan untuk membantu pengguna pada berbagai tugas dalam domain yang spesifik, model *chabot* Mistral 7B dapat dengan mudah disesuaikan dalam tugas tertentu. Model ini dapat disesuaikan untuk obrolan dan memiliki kinerja yang telah mengungguli model *chatbot* Llama 2 dengan parameter 13B (Jiang et al., 2023).

Beberapa keunggulan model *chatbot* Mistral 7B dibandingkan model Llama2, yaitu:

1. Mistral 7B mengungguli Llama 2 13B di semua *benchmark*.
2. Mistral 7B mengungguli Llama 1 34B di banyak tolok ukur.
3. Mistral 7B mendekati kinerja CodeLlama 7B pada kode, namun tetap bagus dalam tugas-tugas bahasa Inggris.
4. Mistral 7B menggunakan *Grouped-query attention* (GQA) untuk inferensi yang lebih cepat.
5. Mistral 7B menggunakan *Sliding Window Attention* (SWA) untuk menangani urutan yang lebih panjang dengan biaya yang lebih kecil. Mistral 7B memiliki lisensi Apache 2.0 dan dapat digunakan tanpa batasan.

Mistral 7B telah diuji pada semua tolok ukur dengan jalur evaluasi untuk mengukur kinerja pada berbagai tugas dan kategori. Semua model dievaluasi ulang pada semua metrik dengan jalur evaluasi yang sama untuk mendapat perbandingan yang akurat. Pada hasil pengujian menunjukan bahwa Mistral 7B mengungguli Llama 2 7B dan Llama 2 13B pada semua benchmark. Model Mistral 7B jauh lebih unggul dari Llama 1 34B dalam matematika, pembuatan kode, dan tolok ukur penalaran. Performa secara detail model Mistral 7B dengan model Llama 2 di beberapa versi dalam menjalankan semua evaluasi model dapat dilihat pada

Gambar 2.1. Mistral 7B mengungguli Llama 2 13B di semua metrik.



Gambar 2.1. Perbandingan performa pada Model Mistral 7B dengan Llama2 (Sumber: Jiang et al., 2023)

## *Model Llama 2*

Llama 2 merupakan kumpulan model bahasa besar yang telah dilatih dan disempurnakan dengan skala mulai dari 7 miliar hingga 70 miliar parameter. Llama 2-Chat merupakan LLM yang telah disempurnakan dari beberapa versi model sebelumnya yang dapat digunakan untuk penelitian dan telah merilis model ke masyarakat umum untuk penelitian dan penggunaan komersial, Peneliti pada jurnal ini telah merilis beberapa model diantara:

1. Llama 2, versi terbaru dari Llama 1, telah dilatih menggunakan campuran data baru yang tersedia untuk umum serta meningkatkan ukuran korpus pada pra-pelatihan sebesar 40%, menggandakan panjang konteks model, dan mengadopsi perhatian kueri yang dikelompokkan (Ainslie et al., 2023). Peneliti pada jurnal ini telah merilis varian Llama 2 dengan parameter 7B, 13B, dan 70B.

2. Llama 2-Chat adalah versi Llama 2 yang disempurnakan dan dioptimalkan untuk kasus penggunaan dialog. Peneliti pada jurnal ini merilis varian model Llama 2-Chat dengan parameter 7B, 13B, dan 70B juga.

Model Llama 2 dapat digunakan untuk umum, penelitian dan penggunaan komersial, serta telah dioptimalkan untuk kasus penggunaan dialog. Model Llama 2 mengungguli model obrolan sumber terbuka pada sebagian besar tolok ukur yang telah diuji, dan berdasarkan evaluasi oleh manusia terhadap kegunaan dan keamanan, model Llama 2 dapat menjadi alternatif pengganti yang cocok untuk model sumber tertutup. Peneliti pada jurnal ini memberikan penjelasan rinci tentang pendekatan yang dilakukan pada proses penyempurnaan dalam peningkatan kinerja Llama 2-Chat sehingga memungkinkan komunitas mengembangkan model Llama 2 serta dapat berkontribusi pada pengembangan LLM yang bertanggung jawab (Touvron et al., 2023).

## *Fine-tuning*

*Fine-tuning* artinya *training model* dengan tugas yang lebih spesifik menggunakan suatu model yang telah dilatih dengan dataset yang besar. Biasanya untuk *training model* menggunakan metode *unsupervised learning*.



Gambar 2.2. Alur *Fine-tuning* model (Sumber: Fatyanosa, 2020)

*Fine-tuning* dilakukan dengan menggunakan *pre-trained language model* seperti BERT dan GPT yang telah dilatih menggunakan dataset yang sangat besar. Model yang dihasilkan mampu memahami model bahasa yang dilatih. Jika tidak menggunakan *pre-trained language model*, maka diperlukan *training* dari awal yang pasti membutuhkan banyak waktu dan *resource*. Jika menggunakan *pre-trained language model*, maka dengan mudah dapat digunakan model yang telah dilatih sebelumnya dan cukup dilakukan *fine-tuning* sesuai dengan dataset atau tugas yang lebih spesifik. Dengan kata lain dengan menggunakan *pre-trained language model*, maka dapat lebih cepat untuk mendapatkan hasil tanpa harus dilakukan pelatihan dari awal. *Fine-tuning* membutuhkan data yang lebih sedikit dan biasanya menggunakan metode *supervised learning* yang sangat terbatas jumlahnya. Jika *deep learning* dilatih dengan data yang sedikit, ada kemungkinan dihasilkan performa yang rendah namun dikarenakan *pre-trained language model* sudah dilatih dengan dataset yang besar, maka saat *fine-tuning* hanya diperlukan dataset dengan jumlah yang lebih sedikit. Model dengan *fine-tuning* memiliki performa yang lebih baik, sebelum muncul *pre-trained language model*, arsitektur model dibuat dengan sedemikian rupa sehingga dapat menghasilkan performa yang lebih tinggi dalam tugas tertentu. Namun saat ini, model-model tersebut telah dikalahkan oleh *pre-trained language model* seperti BERT dan GPT di berbagai tugas yang spesifik dalam NLP (Fatyanosa, 2020).

## *Retrieval-Augmented Generation*

*Retrieval-augmented generation* (RAG) merupakan suatu pendekatan pada model AI yang dapat menghasilkan teks dengan menggabungkan kemampuan *retrieval* (pengambilan informasi) dan *generation* (pembangkitan teks) (Wicaksono, 2023).

Beberapa fitur utama dari *retrieval-augmented generation*:

1. Model dapat mengambil informasi yang relevan dari berbagai sumber data eksternal (seperti basis data fakta, artikel, dan lainnya) yang dapat memperkaya konten yang dihasilkan. Hal ini dapat memungkinkan model dalam menghasilkan teks dengan informasi akurat dan *up-to-date*.

2. Informasi yang diambil kemudian digunakan dalam membimbing pada proses *generate* teks sehingga keluaran yang dihasilkan dapat relevan dengan topik yang diminta dan berisi fakta-fakta yang akurat.

3. Pada umumnya selain mengambil informasi faktual, metode *retrieval-augmented* *generation* juga mengambil contoh teks relevan dari kumpulan data, contoh teks tersebut kemudian digunakan model sebagai bahan acuan untuk menghasilkan teks baru, sehingga gaya bahasa dan struktur kalimat menjadi lebih natural.

4. Model dapat menghasilkan teks dengan kualitas yang lebih baik dan akurat dibandingkan model generatif teks murni, karena didukung dengan komponen *retrieval* di dalam sistemnya.

Jadi secara ringkas, *retrieval-augmented generation* merupakan penggabungan kemampuan dalam pengambilan informasi dan menghasilkan keluaran dalam bentuk teks dengan kualitas yang tinggi. Dengan RAG, model AI tidak hanya mengandalkan pengetahuan yang sudah ada di dalam dirinya, tetapi juga bisa mengambil informasi terbaru dari sumber-sumber eksternal seperti basis data Universitas, kemudian informasi tersebut digunakan untuk membimbing proses *generate* jawaban.

Manfaat lain dengan penerapan RAG antara lain dapat meningkatkan akurasi, kemutakhiran dan kesesuaian jawaban terhadap konteks pertanyaan pengguna. Selain itu, gaya bahasa jawaban juga bisa disesuaikan agar lebih natural (Wicaksono, 2023). Arsitektur pada *Retrieval augmented generation* (RAG) dapat dilihat pada Gambar 2.3.



Gambar 2.3. Arsitektur pada *Retrieval-augmented generation* ([RAG](https://www.researchgate.net/figure/Retrieval-Augmented-Generation-Architecture_fig1_378364457))

(Sumber: Rakotoson, Loïc et al., 2024)

## Kuantisasi

Istilah kuantisasi dapat mengacu pada proses pemetaan nilai tak terhingga yang kontinu pada sekumpulan nilai terhingga yang lebih kecil. Dalam konteks LLM, istilah ini mengacu pada proses konversi bobot model dari tipe data presisi tinggi ke tipe data dengan presisi yang rendah.

1. Presisi Jaringan Syaraf

LLM pada dasarnya merupakan jaringan saraf, model komputasi yang direpresentasikan dalam memori GPU atau RAM sebagai *Tensor - array* angka multidimensi. Untuk menyimpan model ini dapat menggunakan berbagai jenis seperti *Float64*, *Float16*, atau bahkan bilangan bulat. Jenis data yang dipilih akan mempengaruhi pada jumlah "digit" akan digunakan dalam memori dan berpengaruh pada ukuran memori. Ukuran variabel disebut sebagai presisi yang menunjukkan berapa banyak "digit" yang digunakan dalam merepresentasikannya di dalam memori.



Gambar 2.4. Tipe data LLM terkuantisasi (Sumber: Neves, 2023)

Secara umum, untuk kasus penggunaan presisi tinggi di dalam jaringan neural dikaitkan dengan akurasi yang lebih baik dan pelatihan yang lebih stabil. Penggunaan presisi tinggi juga lebih mahal secara komputasi karena memerlukan lebih banyak perangkat keras dan membutuhkan perangkat keras yang lebih canggih dan mahal. Penelitian yang sebagian besar dilakukan oleh Google dan NVIDIA mengenai kemungkinan untuk menggunakan presisi yang lebih rendah untuk beberapa operasi jaringan neural menunjukkan bahwa presisi yang lebih rendah dapat dimanfaatkan untuk beberapa operasi pelatihan dan inferensi.

Selain untuk penelitian, kedua perusahaan mengembangkan perangkat keras dan kerangka kerja untuk mendukung operasi dengan presisi rendah. Misalnya, akselerator NVIDIA T4 adalah GPU dengan presisi rendah dan menggunakan teknologi *Tensor Cores* yang secara signifikan lebih efisien daripada K80. TPU Google memperkenalkan konsep *bfloat16*, tipe data primitif khusus yang dioptimalkan untuk jaringan saraf. Ide yang mendasar di balik presisi rendah adalah bahwa jaringan saraf tidak selalu perlu menggunakan semua rentang yang di *float64* bit agar dapat bekerja dengan baik.



Gambar 2.5. Format numerik *bfloat16* (Sumber: Neves, 2023)

Seiring dengan semakin besarnya jaringan saraf, pentingnya memanfaatkan presisi yang lebih rendah berdampak signifikan pada kemampuan dalam menggunakannya. Dengan LLM, hal ini menjadi lebih penting lagi. Sebagai referensi, GPU A100 buatan NVIDIA memiliki memori 80 GB dalam versi tercanggihnya. Pada Tabel 2.1 di bawah ini, dapat dilihat bahwa model LLama2-70B memerlukan memori sekitar 138 GB, yang berarti bahwa untuk menghostingnya, memerlukan beberapa A100. Mendistribusikan model ke beberapa GPU berarti membayar lebih banyak GPU serta infrastruktur *overhead*. Di sisi lain, versi terkuantisasi memerlukan sekitar 40 GB memori, oleh karena itu dapat dengan mudah masuk ke dalam satu A100, sehingga mengurangi biaya inferensi secara signifikan. Contoh ini bahkan tidak menyebutkan fakta bahwa dalam satu A100, penggunaan model terkuantisasi akan menghasilkan eksekusi yang lebih cepat dari sebagian besar operasi komputasi individual.

Tabel 2.1. Kuantisasi 4-bit menggunakan llama.cpp (Sumber: Neves, 2023)

|  |  |  |
| --- | --- | --- |
| Model | Ukuran Asli (FP16) | Ukuran Terkuantisasi (INT4) |
| Llama2-7B | 13,5 GB | 3,9 GB |
| Llama2-13B | 26,1 GB | 7,3 GB |
| Llama2-70B | 138 GB | 40,7 GB |

1. Kuantisasi untuk mengurangi ukuran model

Kuantisasi secara signifikan dapat mengurangi ukuran model dengan mengurangi jumlah bit yang diperlukan pada setiap bobot model. Skenario yang umum adalah pengurangan bobot dari FP16 (*Floating-point 16-bit*) menjadi INT4 (*Integer 4-bit*). Hal ini memungkinkan model untuk berjalan pada perangkat keras yang lebih murah dan/atau dengan kecepatan yang lebih tinggi. Dengan mengurangi presisi bobot, kualitas keseluruhan LLM juga dapat mengalami beberapa dampak.

Pada beberapa studi menunjukkan bahwa dampak ini dapat bervariasi tergantung pada teknik yang digunakan dan model yang lebih besar tidak terlalu berpengaruh pada perubahan presisi. Model yang lebih besar (lebih dari ~70B) mampu mempertahankan kapasitasnya bahkan saat dikonversi ke 4-bit, dengan beberapa teknik seperti NF4 yang menunjukkan tidak ada dampak pada kinerjanya. Oleh karena itu, 4-bit tampaknya menjadi metode terbaik dilihat dari kinerja dan ukuran/kecepatan pada model yang lebih besar ini, sementara 6 atau 8-bit dapat menjadi model yang lebih baik untuk model yang lebih kecil (Neves, 2023).

1. Dua Jenis Kuantisasi pada LLM (Neves, 2023)

Ada dua teknik cara yang digunakan untuk memperoleh model kuantisasi:

1. ***Post-Training Quantization* (PTQ)**: dengan mengubah bobot pada model yang sudah dilatih menjadi presisi yang lebih rendah tanpa pelatihan ulang. Meskipun mudah dan sederhana untuk diterapkan, PTQ dapat sedikit menurunkan kinerja model karena hilangnya presisi dalam nilai bobot.
2. ***Quantization-Aware Training* (QAT)**: Tidak seperti PTQ, QAT mengintegrasikan proses konversi bobot selama tahap pelatihan. Hal ini sering kali menghasilkan kinerja model yang unggul, tetapi lebih menuntut komputasi. Teknik QAT yang banyak digunakan adalah QLoRA.

## GPTQ: *Post-Training Quantization For GPT*

GPTQ merupakan metode kuantisasi pasca pelatihan (PTQ) dengan kuantisasi 4-bit yang berfokus terutama pada inferensi dan kinerja GPU. Ide di balik metode ini yaitu dengan mencoba mengompresi semua bobot menjadi kuantisasi 4-bit dan meminimalkan kesalahan kuadrat rata-rata terhadap bobot tersebut. Selama inferensi, model akan secara dinamis mendekuantisasi bobotnya menjadi ukuran *float16* untuk meningkatkan kinerja sekaligus menjaga memori tetap rendah (Grootendorst, 2023).

## GGUF: *GPT-Generated Unified Format*

GPTQ dapat melakukan kompresi dengan baik menggunakan GPU tetapi akan menjadi kerugian jika tidak memiliki perangkat keras tersebut untuk menjalankannya. GGUF yang sebelumnya GGML merupakan metode kuantisasi yang dapat memungkinkan pengguna untuk menggunakan CPU untuk menjalankan LLM dengan memindahkan beberapa lapisannya ke GPU untuk mempercepatnya. Penggunaan CPU umumnya lebih lambat dibandingkan dengan menggunakan GPU untuk inferensi, metode ini adalah format yang luar biasa bagi pengguna yang menjalankan model pada CPU atau perangkat Apple. Terutama dengan melihat pada model yang lebih kecil dan lebih mumpuni yang bermunculan, seperti Mistral 7B, format GGUF akan tetap ada (Grootendorst, 2023).

## *Dataset*

Dataset merupakan kumpulan data yang diatur dalam format yang terstruktur, seperti tabel atau file, dan berisi informasi dari berbagai sumber. Dataset dapat berupa data numerik, teks, gambar, atau gabungan dari semuanya. Keberagaman jenis data ini memungkinkan penggunaan dalam aplikasi di berbagai industri dan disiplin ilmu.

Dataset dikumpulkan oleh profesional di bidang data, seperti *data analyst*. Untuk bisa digunakan, dataset perlu melewati beberapa tahapan pengolahan data, seperti *data cleaning* dan pemberian kategori/label. Sehingga, dataset yang digunakan biasanya terkumpul berdasarkan kategorinya masing-masing, dan di dalamnya terdapat variabel-variabel yang saling berhubungan (Risyad, 2023).

### Jenis-jenis Dataset

Data dikumpulkan dari berbagai jenis dan berbagai bentuk dapat dikategorikan ke dalam beberapa jenis. Seorang *data analyst* diharuskan memahami perbedaan karakter dari jenis-jenis dataset. Berbagai jenis dataset ini memiliki fungsi dan tujuan yang berbeda. Risyad (2023) mengkategorikan dataset ke dalam beberapa jenis dataset.

1. *Numerical Dataset*

*Numerical dataset* merupakan jenis dataset yang hanya terdiri dari data numerik. Data numerik ini dapat bersifat kuantitatif dan dapat dihitung, diukur, dan diolah dengan menggunakan perhitungan matematis, analisis statistik, atau model prediktif. Contoh kategori pada data numerik meliputi usia, tinggi badan, suhu, pendapatan, dan banyak lagi.

2. *Bivariate Dataset*

*Bivariate dataset* merupakan jenis dataset yang hanya berisi dua variabel yang saling terkait satu sama lain. Dataset jenis ini digunakan untuk menganalisis hubungan atau korelasi antara dua variabel. Contoh kategori *bivariate dataset* seperti data berat dan tinggi badan dari sekelompok individu, dimana setiap entri data berisi berat dan tinggi badan tiap-tiap individu.

3. *Multivariate Dataset*

*Multivariate dataset* merupakan dataset yang berisi lebih dari dua variabel. Data ini mencakup beberapa atribut atau dimensi, dan setiap entri data memiliki beberapa nilai yang saling berkaitan dengan atribut atau dimensi tersebut. Analisis untuk jenis dataset ini membantu mengidentifikasi ketergantungan antara banyak variabel. Contoh *multivariate dataset* seperti data demografi yang mencakup usia, pendapatan, tingkat pendidikan, pekerjaan, dan lain-lain dari sejumlah individu.

4. *Categorical Dataset*

*Categorical dataset* merupakan suatu jenis dataset yang berisi data dalam bentuk kategori atau kelompok. Dataset ini bersifat kualitatif dan tidak dapat diukur dengan angka. Contoh *data kategorikal* meliputi jenis kelamin (laki-laki/perempuan), status pernikahan (menikah/belum, menikah/cerai), warna mata (cokelat/biru/hijau), dan lain sebagainya.

5. *Correlation Dataset*

*Correlation* *dataset* merupakan suatu jenis dataset yang berisi nilai-nilai korelasi dari dua atau lebih variabel dalam bentuk matriks atau tabel. Nilai korelasi ini menggambarkan sejauh mana hubungan linier antara variabel-variabel tersebut. Nilai korelasi berkisar dari -1 hingga +1, dengan nilai -1 menunjukkan korelasi negatif sempurna, 0 menunjukkan tidak adanya korelasi, dan +1 menunjukkan korelasi positif sempurna. Korelasi negatif menandakan bahwa kedua variabel memiliki hubungan yang berbanding terbalik, sedangkan korelasi positif menandakan kedua variabel bergerak ke arah yang sama (Risyad, 2023).

## USK Dataset

Untuk pelatihan *chatbot* yang diusulkan dalam penelitian ini, dataset ini terbagi 2 yang pertama sebanyak 20 dataset Q&A yang akan digunakan untuk *fine-tuning* dan sebanyak 231 pernyataan diambil dari area "pertanyaan yang sering diajukan oleh mahasiswa baru di USK" di bagian Unit Layanan Terpadu (ULT) Universitas Syiah Kuala (ult.usk.ac.id, 2024) dan disimpan ke dalam format .pdf untuk digunakan dengan metode RAG. Dataset ini secara manual dibagi menjadi beberapa topik berdasarkan topik terkait akademik di USK. Dataset ini terbagi ke dalam beberapa topik seperti uang kuliah (UKTB), akademik, umum, beasiswa, dan lain-lain.

## *Python*

Bahasa pemrograman yang bersifat *interpreter* yaitu Python merupakan salah satu bahasa interaktif dengan *object-oriented* yang beroperasi hampir di semua *platform*, seperti pada Windows, Mac, Linux dan lainnya, dengan sintaksnya yang cukup jelas dan elegan *python* dikategorikan sebagai salah satu bahasa pemrograman yang dapat dipelajari dengan mudah, modul - modul pada *python* dikombinasikan dengan menggunakan struktur data tingkat tinggi yang efisien.

Bahasa pemrograman *python* sendiri diciptakan di Belanda pada tahun 1990 oleh Guido Van Rossum dan Namanya diambil dari acara televisi kesukaan Guido yaitu *Monty Python’s Flying Circus*. Secara luas python digunakan sebagai bahasa pemrograman yang dipakai dalam industri dan juga dalam Pendidikan dikarenakan sintaks yang intuitif, ringkas, sederhana dan terdapat pustaka yang banyak.

*Python* juga merupakan pemrograman yang berorientasi pada objek atau OOP. Kelas (*class*) dalam python merupakan sebuah objek berupa data dan merupakan suatu alat yang dapat digunakan dalam pengembangan perangkat lunak yang nantinya dapat digunakan kembali. *Python* merupakan bahasa yang memiliki kemampuan, menggabungkan kapabilitas, dan sintaksis kode dengan sangat jelas, dan juga dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif. Walaupun *Python* tergolong bahasa pemrograman dengan level tinggi, nyatanya *Python* dirancang sedemikian rupa agar mudah dipelajari dan dipahami. *Python* memiliki tata bahasa dan *script* yang sangat mudah untuk dipelajari. *Python* juga memiliki sistem pengelolaan data dan memori otomatis. Selain itu modul pada *Python* selalu diupdate (diskominfo.kedirikab.go.id, 2023).

# BAB III METODOLOGI PENELITIAN

## Tempat dan Waktu Penelitian

Penelitian dilakukan di Prodi Magister Kecerdasan Buatan Departemen Informatika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Syiah Kuala. Selama kurang lebih 6 (enam) bulan, peneliti akan melakukan penelitian yang intensif dengan fokus pada topik yang telah ditentukan yang dimulai dari bulan April sampai dengan September 2024.

## Alat dan Bahan

Alat dan bahan yang digunakan pada saat kegiatan penelitian ditunjukkan pada

Tabel 3.1 Alat dan Bahan.

Tabel 3.1. Alat dan Bahan

|  |  |
| --- | --- |
| No | Alat dan Bahan |
| 1. | Komputer dengan spesifikasi yang cukup untuk menjalankan Google Colab |
| 2. | Google Colab dengan NVIDIA Tesla T4 GPU |
| 3. | Dataset informasi penerimaan mahasiswa baru di Universitas Syiah Kuala |
| 4. | Model Mistral 7B |

## Metode Penelitian

Penelitian terdiri dari tiga tahapan utama. Tahapan pertama adalah mengumpulkan dataset berupa pernyataan terkait akademik di USK. Metode yang akan digunakan untuk mengumpulkan dataset ini yaitu dengan meminta data pertanyaan akademik ke bagian layanan terpadu USK serta melakukan *preprocessing* pada dataset. Selanjutnya tahap kedua adalah melakukan pengembangan menggunakan dataset yang sudah terkumpul pada model Mistral 7B dengan metode *fine-tuning* dan RAG dengan menggunakan Google Colab. Kemudian terakhir pada tahap ketiga menguji model Mistral 7B dengan mengajukan beberapa pertanyaan dasar mengenai akademik di USK dan mengevaluasi kinerja pada metode *fine-tuning* dan RAG. *Flowchart* untuk alur peneliti dapat dilihat pada Gambar 3.1.



Gambar 3.1. *Flowchart* Alur Penelitian

### Pengumpulan dan Persiapan Dataset

Tahap awal adalah mengumpulkan data mentah dari Unit Layanan Terpadu USK, Humas USK dan berbagai website USK yang mencakup berbagai jenis data mengenai proses administrasi dan akademik di USK. Proses pengumpulan data terdiri dari informasi akademik (misalnya terkait pembayaran UKTB, pengurusan KTM, dan juga Beasiswa). Data tersebut kemudian akan dilakukan *preprocessing* dengan mengubah ke dalam 2 format pertama ke dalam formatinstruksi dengan eksentensi .csv dan kedua ke dalam format .pdf yang akan digunakan untuk mengembangkan LLM dengan domain terkait akademik di USK.

### *Fine-tuning* LLM

*Fine-tuning* dilakukan untuk mengkhususkan LLM pada data pelatihan dengan tugas pembelajaran tertentu. Artinya, perlu dilakukan identifikasi terhadap data masukkan, tujuan pembelajaran, dan proses pelatihan dengan benar dan tepat. *Fine-tuning* akan memerlukan alur pelatihan, alur penyajian, dan kemungkinan alur pelatihan berkelanjutan (Benveniste, 2023). Alur pelatihan dapat dilihat pada Gambar 3.2.



Gambar 3.2. Alur Pelatihan pada LLM (Sumber: Benveniste, 2023)

1. **Registri model**: Registri model ini berisi model asli atau dapat berupa versi berbeda dari model yang telah disempurnakan.
2. **Modul kuantisasi**: Mengkuantisasi model saat ini digunakan untuk menghemat biaya dan penggunaan memori, tetapi hal ini bukanlah suatu keharusan. Model dikuantisasi yang dihasilkan perlu disimpan dalam registri model. Kuantisasi berarti mengkonversi bobot model dari *float* menjadi bilangan bulat. Operasi ini membagi ukuran model dengan 4.
3. **Penyimpanan fitur**: Data yang digunakan untuk pelatihan pada model bahasa besar dipersiapkan dengan cermat dan disimpan di dalam penyimpanan fitur.
4. **Modul validasi dan pra-pemrosesan data**: Data yang dimasukkan ke dalam alur pelatihan perlu dilakukan validasi dan kemungkinan besar diproses terlebih dahulu sebelum dilakukan pelatihan.
5. **Modifikasi LoRA / QLoRA**: Proses ini merupakan penyempurnaan LLM menggunakan Adaptor Tingkat Rendah (LoRA) atau versi Terkuantisasinya (QLoRA). Idenya yaitu dengan mengganti bagian di dalam model menjadi beberapa matriks besar dengan matriks yang lebih kecil pada perhitungan gradien. Pada saat proses penyempurnaan ini hanya dilakukan dengan memperbarui bobot matriks yang baru. Matriks gradien ini jauh lebih kecil sehingga hanya membutuhkan memori GPU yang lebih sedikit. Bobot yang telah dilatih sebelumnya akan dibekukan, sehingga tidak perlu menghitung gradien pada sebagian besar parameter.
6. **Registri LoRA / QLoRA**: LoRA / QloRA merupakan registri pada model untuk melacak bobot yang telah disesuaikan dan versinya.
7. **Modul validasi model**: Model yang telah dilatih perlu divalidasi menggunakan data validasi. Hal ini dilakukan agar model dapat memberikan jawaban yang sesuai dengan domain pada tugas tersebut. Hal ini menjadi rumit karena tidak hanya menyempurnakan model pada tugas tertentu, namun juga harus tetap mempertahankan kemampuan aslinya. Untuk tugas tersebut dengan memiliki data validasi yang tepat, dapat menghindari kehilangan data pada model aslinya sehingga tidak melupakan informasi yang ada pada model sebelumnya.

Pada Gambar 3.2 dapat dilihat alur *fine-tuning* dimulai dari Registri model dengan menggunakan model asli kemudian model tersebut dikuantisasi dengan menggunakan metode QLoRA. Penyimpanan fitur berisi kumpulan data yang digunakan untuk *fine-tuning,* data ini dilakukan validasi dan pra-pemrosesan sebelum pelatihan. Pada tahap selanjutnya LLM disempurnakan menggunakan Adaptor Tingkat Rendah (LoRA) menggunakan versi Terkuantisasinya (QLoRA). Validasi model dilakukan setelah menyempurnakan model dengan mempertahankan kemampuan aslinya. Setelah model siap di *fine-tuning* dan didapatkan model dengan bobot QLoRA selanjutnya dilakukan serangkaian pengujian untuk memastikan model sesuai dengan alur kerja yang tersedia dan dengan eksperimen pengujian untuk mengujinya terhadap model produksi. Setelah itu model dapat diusulkan menjadi model baru sebagai model produksi.

### LLM dengan metode RAG

*Retrieval Augmented Generation* atau RAG berarti mengekspos LLM ke data baru yang disimpan dalam database. Tanpa mengubah LLM, dengan memberikan konteks data tambahan agar LLM dapat menjawab pertanyaan dengan informasi mengenai subjek tertentu. Ide RAG adalah untuk mengkodekan data yang ingin diekspos pada LLM ke dalam *embeddings* dan mengindeks data tersebut ke dalam database vektor. Saat pengguna mengajukan pertanyaan, pertanyaan tersebut diubah menjadi penyematan, dan dapat digunakan untuk mencari penyematan serupa di dalam database vektor. Setelah menemukan penyematan serupa, kemudian akan dibuatkan *prompt* dengan data terkait untuk memberikan konteks bagi LLM untuk menjawab pertanyaan tersebut. Kemiripan diukur menggunakan metrik kesamaan kosinus (Benveniste, 2023). Alur pada RAG dapat dilihat pada Gambar 3.3.



Gambar 3.3. Alur Pada RAG (Sumber: Benveniste, 2023)

**Sistem Pengindeksan pada RAG**

RAG tidak memiliki jalur pelatihan hanya memerlukan *pipeline* pengindeksan dan *pipeline* penyajian. Pipa pengindeksan digunakan untuk mengubah data menjadi representasi vektor dan mengindeksnya dalam database vektor seperti yang dapat dilihat pada Gambar 3.4.



Gambar 3.4. Pipa pengideksan pada RAG (Sumber: Benveniste, 2023)

### Evaluasi Kinerja Model

Metrik yang digunakan untuk evaluasi kinerja model pada penelitian ini yaitu Metrik Rouge. Metrik Rouge digunakan untuk mengevaluasi model pada tugas NLP sehingga dapat dibandingkan ringkasan teks yang dihasilkan oleh model dengan ringkasan pada referensi.

***Recall-Oriented Understudy for Gisting Evaluation* (ROUGE)**

Metrik Rouge merupakan metrik evaluasi yang digunakan dalam tugas NLP untuk membandingkan ringkasan teks yang dihasilkan komputer dengan ringkasan referensi (yang dibuat oleh manusia) (Lin, 2004). Rouge sebagian besar digunakan untuk mengevaluasi tugas peringkasan teks. Nilai metrik Rouge berkisar antara 0 hingga 1. 1 adalah skor yang tertinggi, menunjukkan bahwa ringkasan yang dihasilkan komputer dan ringkasan referensi memiliki kesamaan yang tinggi. Rouge-1, Rouge-2, dan Rouge-L membandingkan dua ringkasan dengan rincian yang berbeda (Amanat, 2024).

* + - 1. Rouge-1

Rouge-1 mengukur ketepatan unigram (kata tunggal) yang tumpang tindih antara teks yang dihasilkan dan teks referensi (buatan manusia).

|  |  |
| --- | --- |
|  | (3.1) |
|  | (3.2) |
|  | (3.3) |

* + - 1. Rouge-2

Rouge-2 mengukur ketepatan bigram yang tumpang tindih antara teks yang dihasilkan dan teks referensi (buatan manusia). Rumus pada Rouge-2 sama dengan Rouge-1, namun yang digunakan sepasang kata yaitu bigram, bukan unigram. Bigram mengkompensasi masalah posisi kata Rouge-1 sampai batas tertentu.

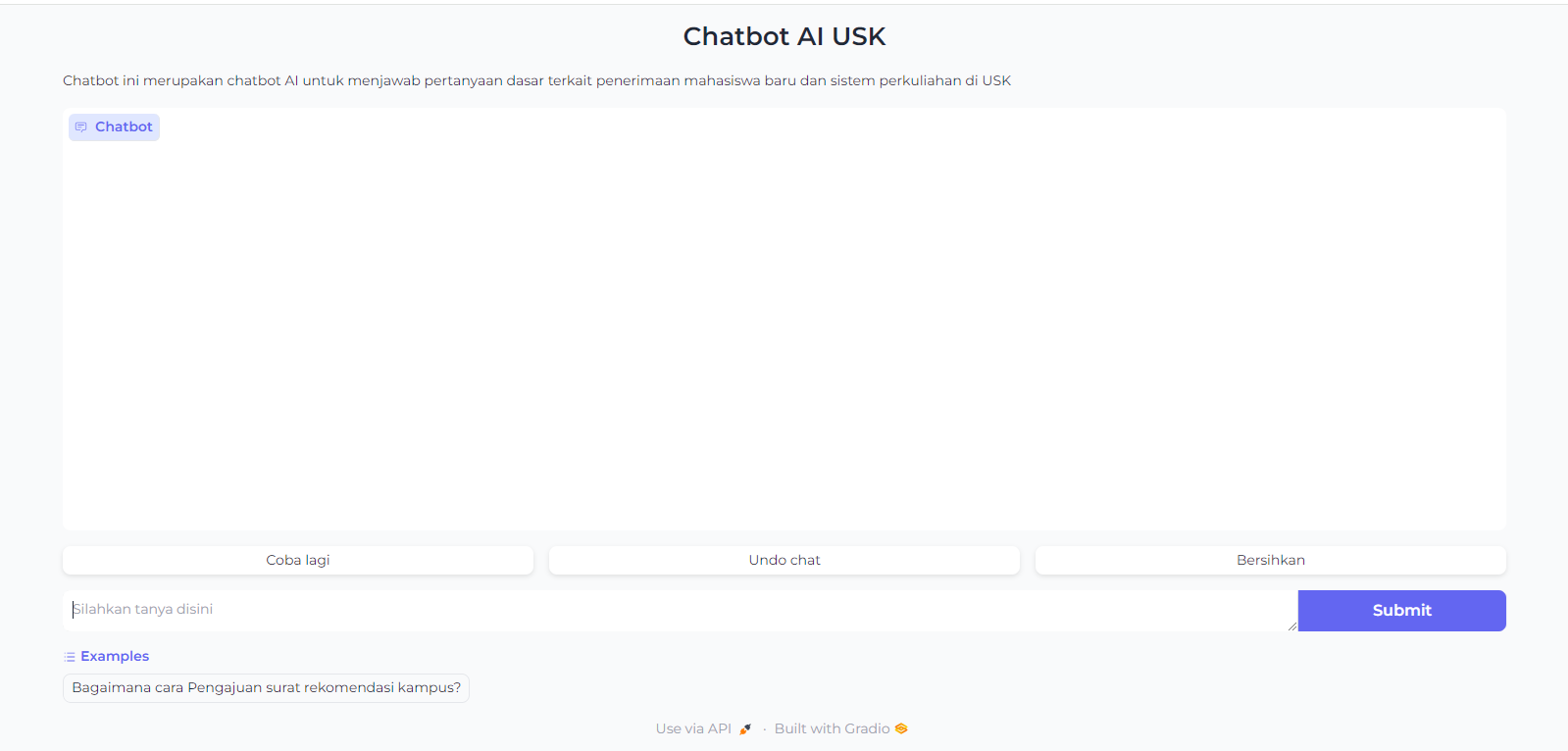
* + - 1. Rouge-L

Tidak seperti Rouge-1, dan Rouge-2, Rouge-L tidak melihat ke dalam unigram atau bigram, melainkan kesesuaian dengan LCS (*Longest Common Subsequence*) atau urutan kata terpanjang dalam referensi dan teks yang dihasilkan manusia.

|  |  |
| --- | --- |
|  | (3.4) |
|  | (3.5) |
|  | (3.6) |

### Perancangan dan Pengembangan Sistem UI

Pada tahap ini, model yang telah dilatih kemudian dilakukan perancangan antarmuka pengguna pada LLM sehingga dapat dihasilkan aplikasi interaktif yang ramah bagi pengguna. Dengan memanfaatkan pustaka *Python* sumber terbuka yang dapat digunakan untuk membuat demo aplikasi berbasis web. Pustaka *Python* sumber terbuka yang digunakan, yaitu *gradio* yang dapat dengan cepat membuat antarmuka pengguna yang indah. Antarmuka pada *gradio* dapat disesuaikan dengan pengaturan pada judul, deskripsi, dan tampilan. Saat diluncurkan, antarmuka web memungkinkan pengguna untuk berinteraksi dengan chatbot, cara kerjanya cukup dengan mengetikkan pertanyaan pada kolom chat dan chatbot akan menghasilkan inferensi jawaban berdasarkan pertanyaan yang diajukan oleh pengguna (Khan, 2023).



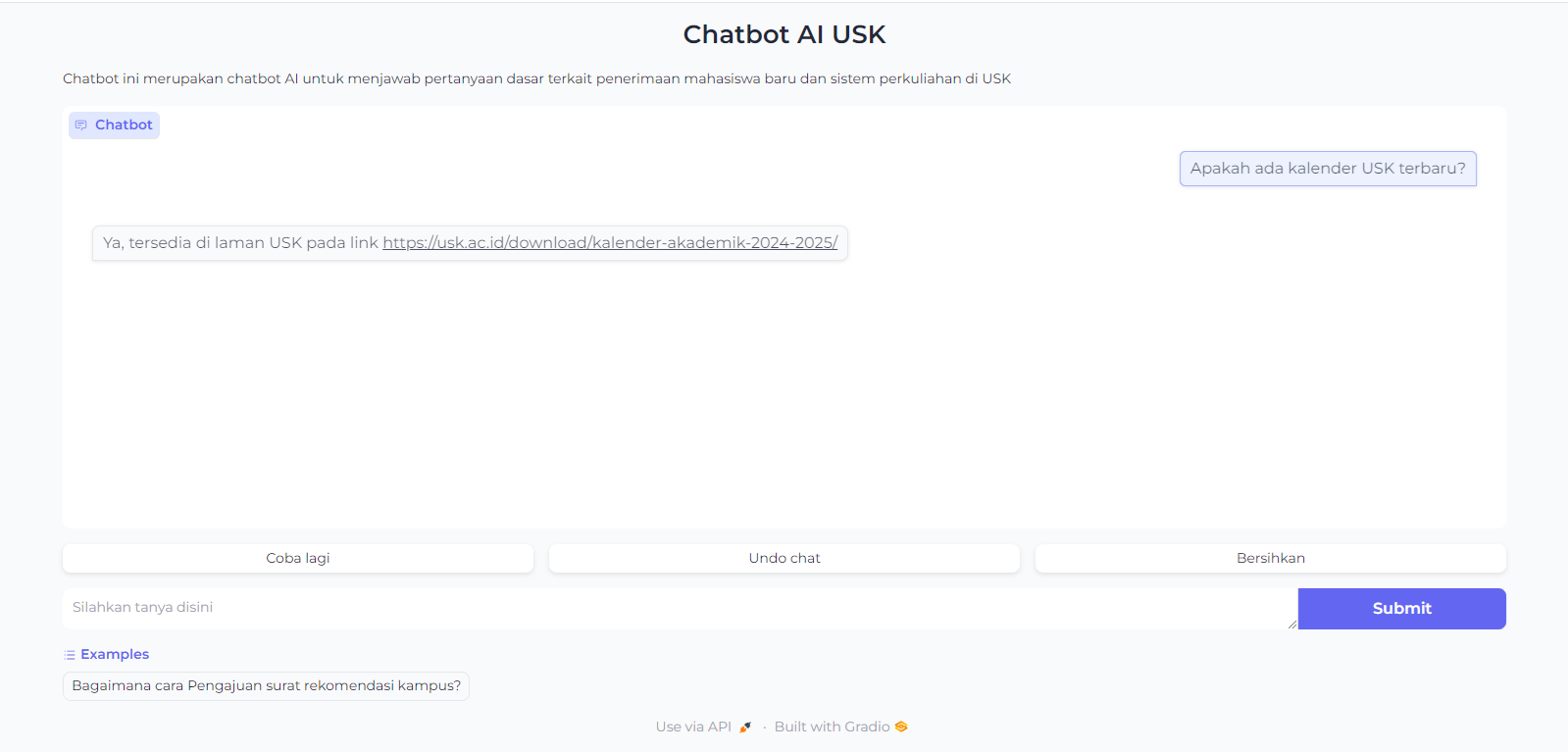
Gambar 3.5. Desain UI Chatbot menggunakan Gradio (www.gradio.app, 2024)

# BAB IV HASIL DAN PEMBAHASAN

Luaran yang dihasilkan dalam penelitian ini adalah sebuah aplikasi *chatbot* berbasis web dengan memanfaatkan *Large Language Model* (LLM) Mistral 7B yang akan digunakan sebagai alternatif dalam membantu pengguna mendapatkan informasi di USK. *Chatbot* ini dikembangkan dengan menggunakan kumpulan data dengan memanfaatkan kemampuan peringkasan teks LLM untuk menghasilkan teks informasi terkait akademik di USK. Ada beberapa tahapan yang dilakukan dalam memanfaatkan dan mengembangkan LLM, yaitu dengan mengumpulkan data terkait akademik di USK kemudian kumpulan data tersebut dilakukan *preprocesing* dengan mengubah data mentah ke dalam format .csv dan .pdf. Adapun setelah tahap *preprocesing* kemudian dilakukan *fine-tuning* pada model Mistral 7B dan juga untuk memudahkan dalam mengelola data dilakukan RAG dengan konsep *embeddings*. Tujuan RAG, yaitu untuk mengatasi batasan pada *AI generatif* karena setiap kali ada pertanyaan yang memerlukan informasi yang berada di luar korpus pelatihan LLM, akan mengakibatkan halusinasi, ketidakakuratan, atau pembelokan pada teks yang dihasilkan. Adapun tahap berikutnya menguji dan mengevaluasi teks yang dihasilkan dengan metode Rouge untuk membandingkan teks yang dihasilkan oleh model dengan ringkasan referensi yang diberikan. Tahap terakhir dengan membuat UI atau antarmuka web agar dapat diakses oleh pengguna.

## Tampilan Chatbot dan Hasil Inferensi LLM

Dalam penelitian ini dataset yang digunakan adalah berupa kumpulan data informasi mengenai akademik di USK dengan jumlah dataset yang didapat sebanyak 231 data pertanyaan dan jawaban. Kumpulan data tersebut akan digunakan pada LLM untuk dapat menghasilkan teks yang mirip dengan teks yang dihasilkan manusia melalui proses yang dikenal sebagai inferensi generatif. Hasil inferensi dari model Mistral 7B dapat dilihat pada Gambar 4.1.



Gambar 4.1. Contoh Hasil inferensi Chatbot menggunakan desain Gradio UI

Seperti yang ditunjukan pada gambar dengan memanfaatkan kemampuan peringkasan teks LLM untuk menghasilkan teks dari masukkan informasi yang telah diberikan chatbot ini tidak hanya sebagai alternatif selain web tetapi juga akan membantu menghemat waktu dan tenaga staf admin universitas.

## Kondisi Dataset

Dataset yang digunakan untuk *fine-tuning* LLM Mistral 7B yaitu sebanyak 20 data terkait informasi akademik di USK. Dari dataset yang telah didapatkan kemudian dilakukan *preprocessing* data, disini dataset akan diubah kedalam format instruksi Q&A dan disimpan dalam format berekstensi .csv. Dataset ini juga digunakan pada metode RAG yang dibuat dalam bentuk pernyataan dan disimpan ke dalam format berekstensi .pdf. Format dengan ekstensi .csv disimpan ke dalam repositori *huggingface* dan digunakan pada tahap *finetuning* Model. Data dengan format .pdf didapatkan sebanyak 231 data dan digunakan pada metode RAG dengan tujuan agar data tersebut dapat dikelola dengan lebih mudah. Terdapat 20 sampel dataset Q&A yang digunakan untuk *fine-tuning* pada LLM ditampilkan pada Lampiran 1. Beberapa contoh sampel dari dataset yang digunakan pada metode RAG ditampilkan pada Lampiran 2.

## Kuantisasi dan Pelatihan Model

Pada Gambar 4.2, proses *fine-tuning* model Mistral 7B dapat dibagi menjadi tiga tahap sebagai berikut:

1. *Fine-tuning* Model dan Persiapan Data

*Fine-tuning*

Mistral 7B

Dataset USK

USK Mistral 7B

Gambar 4.2. Proses *Fine-tuning* pada Model Mistral 7B

Proses *fine-tuning* dilakukan dengan bantuan pustaka *Unsloth* yang dapat membuat *fine-tuning* LLM 5x lebih cepat dan menggunakan VRAM 80% lebih sedikit. Model hasil *fine-tuning* selanjutnya dilakukan kuantisai ke dalam format 4bit dengan teknik QLoRA. Tujuan kuantisai adalah untuk mengurangi presisi *floating-point* bobot model melalui proses kuantisasi ini model akan dikonversi dari presisi tinggi ke presisi rendah. Dalam proses ini, bobot model di konversi ke format int4 melalui lapisan kuantisasi dan disimpan di GPU. Komputasi inti dilakukan pada CUDA, yang dapat mengurangi konsumsi memori dan meningkatkan efisiensi. Hal ini memungkinkan untuk menyempurnakan model yang lebih besar pada GPU tingkat konsumen. Model hasil *fine-tuning* disimpan dalam repositori *huggingface* dalam format GGUF*.*

1. *Pre-trained* Model

Sebelum tahap pelatihan, dilakukan analisis terhadap LLM terkenal dari kumpulan LLM yang ada saat ini untuk memilih model yang paling sesuai untuk tujuan penelitian ini. Evaluasi pada LLM terutama dengan mempertimbangkan beberapa aspek utama seperti kemampuan beradaptasi domain, kompatibilitas dengan standar di bidang akademik, kemampuan bilingual, ketersediaan sumber terbuka, efisiensi parameter, serta pertimbangan biaya dan perizinan. Pertimbangan ini juga melihat beberapa metrik evaluasi yang telah diuji di berbagai tolok ukur.

1. *Parameter-Efficient Fine-Tuning* (PEFT)

Berdasarkan pertimbangan kumpulan data yang terbatas dan untuk mengurangi biaya pelatihan serta memitigasi potensi risiko, peneliti memilih metode *Progressive Layer Freezing and Fine-tuning* (PEFT) untuk menyempurnakan model Mistral 7B. Metode PEFT saat ini merupakan salah satu pendekatan utama untuk menyempurnakan LLM. PEFT secara selektif menyempurnakan sejumlah kecil parameter model tambahan, sehingga secara signifikan dapat mengurangi biaya komputasi dan penyimpanan, serta memungkinkan adaptasi yang efisien dari LLM yang telah dilatih sebelumnya ke berbagai domain aplikasi.

Pada penelitian ini Secara khusus menggunakan metode penyesuaian tingkat rendah (LoRA) (Hu et al., 2022) untuk menyempurnakan model bahasa besar. LoRA melibatkan pembekuan bobot pra-pelatihan dan pembuatan matriks versi peringkat rendah untuk lapisan kueri dan nilai. Matriks peringkat rendah ini memiliki parameter yang jauh lebih sedikit dibandingkan model aslinya, sehingga memungkinkan penyesuaian dengan penggunaan memori GPU yang lebih sedikit.

Pengaturan *hyperparameter* yang digunakan pada proses *fine-tuning* menggunakan metode QLoRa ditunjukkan pada Tabel. Pada Tabel 4.1 menunjukkan konfigurasi dalam membuat *instance* *FastLanguageModel.from\_pretrained*, menggunakan komponen *unsloth*, dengan konfigurasi spesifik seperti panjang urutan maksimum, tipe data, dan pemuatan 4-bit (V., 2024).

Tabel 4.1. Konfigurasi *FastLanguageModel.from\_pretrained*

|  |  |
| --- | --- |
| **Parameter** | **Nilai** |
| model\_name | mistralai/Mistral-7B-v0.1 |
| max\_seq\_length | 2048 |
| dtype | None |
| load\_in\_4bit | True |

Pada Tabel 4.2 Objek *FastLanguageModel* menyediakan atribut *get\_peft\_model* tempat untuk dapat mengkonfigurasi berbagai parameter untuk penyesuaian, seperti jumlah *attention heads*, modul target, *dropout rate*, *LoRa alpha*, dan banyak lagi. Penggunaan *gradien* *checkpointing* dan teknik canggih lainnya menunjukkan kemampuan *unsloth* untuk mengoptimalkan kinerja model (V., 2024).

Tabel 4.2. Konfigurasi FastLanguageModel.get\_peft\_model

|  |  |
| --- | --- |
| **Parameter** | **Nilai** |
| r | 16 |
| target\_modules | ["q\_proj", "k\_proj", "v\_proj", "o\_proj", "gate\_proj", "up\_proj", "down\_proj",], |
| lora\_alpha | 16 |
| lora\_dropout | 0 |
| bias | None |
| use\_gradient\_checkpointing | "unsloth" |
| random\_state | 3407 |
| use\_rslora | False |
| loftq\_config | None |

Langkah berikutnya pada Tabel 4.3 dan Tabel 4.4 adalah menginisialisasi *Supervised Fine-tuning Trainer* yang membantu proses *fine-tuning*. Dimulai dengan menginisialisasi model, bersama dengan kumpulan data yang harus disempurnakan, *tokenizer* dan semua argumen pelatihan yang diperlukan (kecepatan pembelajaran, langkah maksimum, penurunan bobot, pengoptimalan, dll) (V., 2024).

Tabel 4.3. Konfigurasi *SFTTrainer*

|  |  |
| --- | --- |
| **Parameter** | **Nilai** |
| model | model |
| tokenizer | tokenizer |
| train\_dataset | dataset |
| dataset\_text\_field | "text" |
| max\_seq\_length | 2048 |
| dataset\_num\_proc | 2 |
| packing | False |

Tabel 4.4. Konfigurasi *TrainingArguments*

|  |  |
| --- | --- |
| **Parameter** | **Nilai** |
| per\_device\_train\_batch\_size | 2 |
| gradient\_accumulation\_steps | 4 |
| warmup\_steps | 5 |
| max\_steps | 60 |
| learning\_rate | 2e-4 |
| fp16 | not is\_bfloat16\_supported() |
| bf16 | is\_bfloat16\_supported() |
| logging\_steps | 1 |
| optim | "adamw\_8bit" |
| weight\_decay | 0.01 |
| lr\_scheduler\_type | "linear" |
| seed | 3407 |
| output\_dir | "outputs" |

## Pendekatan LLM dengan metode RAG

Pada tahap ini dilakukan pendekatan pada LLM dengan metode RAG. Pendekatan dengan metode ini bertujuan untuk mengatasi batasan pada AI generatif ketika memerlukan informasi yang berada di luar korpus pelatihan LLM. Dengan metode ini akan menghindari LLM yang akan menghasilkan teks yang tidak akurat, halusinasi, atau pembelokan pada saat memberikan jawaban dari pertanyaan yang diberikan. pada metode RAG data yang digunakan yaitu dokumen eksternal yang berisi informasi terkait akademik di USK dan disimpan dalam format berekstensi .pdf. Kumpulan data informasi tersebut kemudian dilakukan *embedding* untuk mengubah teks menjadi vektor sehingga dapat disimpan ke dalam vektor database. Vektor database yang digunakan pada penelitian ini yaitu FAISS. Model yang akan digunakan pada tahap ini yaitu model yang telah terkuantisasi ke dalam format *GPT-Generated Unified Format* (GGUF) sehingga memungkinkan untuk menggunakan CPU saat menjalankan LLM dengan memindahkan beberapa lapisannya ke GPU yang dapat mempercepat kinerja model dalam menghasilkan teks. Proses kueri pada model dalam pengambilan konteks yang paling relevan dari perintah pengguna dengan metode RAG ditampilkan pada Gambar 4.3



Gambar 4.3. Pipeline RAG, selama fase kueri, mengambil konteks yang paling relevan dari perintah pengguna, meneruskannya ke *Large* *Language Model*

(Sumber: docs.llamaindex.ai, 2023)

## Hasil Pengujian dan Evaluasi Hasil Inferensi

Pada tahap ini dilakukan pengujian chatbot dengan memberikan pertanyaan, hasil inferensi teks yang dihasilkan kemudian dihitung dengan menggunakan metode Rouge skor. Dengan menggunakan metrik Rouge, tanggapan dari chatbot dapat dievaluasi untuk melihat kesamaan antara referensi secara kuantitatif dan jawaban yang dihasilkan oleh chatbot. Beberapa contoh Q&A hasil dengan menggunakan metode *fine-tuning* dapat dilihat pada Tabel 4.5. Hasil perhitungan skor Rouge pada model dengan metode *fine-tuning* ditunjukan pada Tabel 4.6. Beberapa contoh Q&A hasil dengan menggunakan metode RAGdapat dilihat pada Tabel 4.8. Hasil perhitungan skor Rouge pada model dengan metode RAG ditunjukan pada Tabel 4.9.

Tabel 4.5. Hasil dengan metode *fine-tuning*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Pertanyaan | Referensi | Jawaban | Skor ROUGE |
| 1 | Bagaimana cara mendapatkan Sertifikat Akreditasi Universitas? | Untuk sertifikat akreditasi silahkan download disini ya https://lpm.usk. ac.id/ untuk sertifikat akreditasi yang terbaru tidak perlu dilegalisir karena sudah ada QR Code. | Untuk sertifikat akreditasi silahkan download disini ya https://lpm.usk.ac.id/ untuk sertifikat akreditasi yang terbaru tidak perlu dilegalisir karena sudah ada QR Code. | Rouge 1 : 1.0, Rouge 2 : 1.0, Rouge L : 1.0 |
| 2 | Bagaimana cara membayar UKT untuk jalur SMMPTN Barat, dan UKT nya apakah di bayar sekalian dengan IPI nya? | Silahkan langsung bayar ke Bank, dan masalah IPI tidak boleh di cicil kecuali hanya Fakultas Kedokteran saja. | Silahkan langsung bayar ke Bank, dan masalah IPI tidak boleh di cicil kecuali hanya Fakultas Kedokteran saja. | Rouge 1 : 1.0, Rouge 2 : 1.0, Rouge L : 1.0 |
| 3 | Bagaimana cara Cetak KTM? | KTM dapat dicetak di laman berkas-akademik.usk.ac.id. | KTM dapat dicetak di laman berkas-akademik.usk.ac.id. | Rouge 1 : 1.0, Rouge 2 : 1.0, Rouge L : 1.0 |
| 4 | Kapan jadwal sanggahan UKT bagi Mahasiswa? | Untuk data UKT yang sudah dibuat tidak bisa diubah kembali dan data UKT itu tidak menjadi masalah yang penting data KRS benar semua dan mengenai banding UKT nanti akan dibuka oleh bagian Keuangan, silahkan pantau infonya di web usk.ac.id. | Untuk data UKT yang sudah dibuat tidak bisa diubah kembali dan data UKT itu tidak menjadi masalah yang penting data KRS benar semua dan mengenai banding UKT nanti akan dibuka oleh bagian Keuangan, silahkan pantau infonya di web usk.ac.id. | Rouge 1 : 1.0, Rouge 2 : 1.0, Rouge L : 1.0 |
| 5 | Bagaimana cara membuat Surat Keterangan Aktif Kuliah? | Terkait masalah Surat Keterangan Aktif Kuliah silahkan mengurus ke bagian Registrasi Akademik. | Terkait masalah Surat Keterangan Aktif Kuliah silahkan mengurus ke bagian Registrasi Akademik. | Rouge 1 : 1.0,  Rouge 2 : 1.0,  Rouge L : 1.0 |

Tabel 4.6. Nilai skor ROUGE dengan Metode *Fine-tuning*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metode | Jumlah Pertanyaan | Skor ROUGE | | |
| R-1 | R-2 | R-L |
| *Fine-tuning* | 20/20 | 1.0 | 1.0 | 1.0 |

Tabel 4.7. Hasil jawaban dengan Paraphrase pada pertanyaan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Pertanyaan Dasar | Pertanyaan Paraphrase | Referensi | Jawaban |
| 1 | Bagaimana cara mendapatkan Sertifikat Akreditasi Universitas? | Apa langkah-langkah untuk memperoleh Sertifikat Akreditasi Universitas? | Untuk sertifikat akreditasi silakan download disini ya https://lpm.usk.ac.id/, untuk sertifikat akreditasi yang terbaru tidak perlu dilegalisir karena sudah ada QR Code. | Untuk memperoleh sertifikat akreditasi, silakan download disini <https://ban-pt.org/>, di bagian bawah halaman terdapat tombol download sertifikat. |
| 2 | Bagaimana cara membayar UKT untuk jalur SMMPTN Barat, dan UKT nya apakah di bayar sekalian dengan IPI nya? | Bagaimana proses pembayaran UKT untuk jalur SMMPTN Barat, dan apakah UKT dibayar sekaligus dengan IPI-nya? | Silahkan langsung bayar ke bank, dan masalah IPI tidak boleh dicicil kecuali hanya Fakultas Kedokteran saja. | Untuk pembayaran UKT, silakan membayar ke Bank Syariah Indonesia cabang Kota Banda Aceh. Nama Rekening: USK, Nomor Rekening: 9144620111-81-X, Nama Pihak Yang Dikreditkan: Universitas Syiah Kuala. Untuk pembayaran IPI sendiri tidak boleh di cicil kecuali hanya sekali saja pada saat pendaftaran pertama kali. |
| 3 | Bagaimana cara Cetak KTM? | Bagaimana cara mencetak KTM? | KTM dapat dicetak di laman berkas-akademik.usk.ac.id. | Untuk mencetak KTM, Anda dapat membayar sebesar Rp 18.000,- (delapan belas ribu rupiah) per lembar. |
| 4 | Kapan jadwal sanggahan UKT bagi Mahasiswa? | Kapan jadwal pengajuan sanggahan UKT bagi Mahasiswa? | Untuk data UKT yang sudah dibuat tidak bisa diubah kembali dan data UKT itu tidak menjadi masalah yang penting data KRS benar semua dan mengenai banding UKT nanti akan dibuka oleh bagian Keuangan, silakan pantau infonya di web usk.ac.id. | Untuk pengisian UKT dapat menunggu sampai Status Buta Warna berubah warna Hijau. |
| 5 | Bagaimana cara membuat Surat Keterangan Aktif Kuliah? | Apa langkah-langkah untuk membuat Surat Keterangan Aktif Kuliah? | Terkait masalah Surat Keterangan Aktif Kuliah silahkan mengurus ke bagian Registrasi Akademik. | Untuk membuat Surat Keterangan Aktif Kuliah silahkan mengurus ke bagian Registrasi Akademik. |

Tabel 4.8. Hasil dengan metode RAG

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Pertanyaan | Referensi | Jawaban | Skor ROUGE |
| 1 | Bagaimana cara mengurus Surat Rekomendasi Kampus? | Mahasiswa yang ingin membuat Surat Rekomendasi Kampus untuk Beasiswa Pengajuannya dapat dilakukan di bagian Kemahasiswaan. | Surat rekomendasi dapat dilakukan di bagian Kemahasiswaan. | Rouge 1 : 0.6363, Rouge 2 : 0.5, Rouge L : 0.6363 |
| 2 | Apakah ada tata cara mendaftar di USK? | Tata cara mendaftar di USK silahkan Klik link berikut: <https://pmb.usk.ac> .id/berita/ penerimaan-mahasiswa-baru-usk-t-a-2023-2024-melalui-seleksi-talenta-usk-berprestasi. | Silahkan klik link berikut: https://pmb. usk.ac.id/berita/  penerimaan-mahasiswa-baru-usk-t-a-2023-2024-melalui-seleksi -talenta-usk-berprestasi. | Rouge 1 : 0.7924, Rouge 2 : 0.7058, Rouge L : 0.7924 |
| 3 | Dimana kita dapat mendownload Kalender Akademik USK tahun 2024? | Kalender Akademik USK tahun 2024 silahkan di download pada link Berikut: https://usk.ac.id/  download/kalender-akademik-2023-2024/. | silahkan di unduh pada link berikut: [https://usk.ac.id/ download/ kalender-akademik-2023-2024/](https://usk.ac.id/download/%20kalender-akademik-2023-2024/). | Rouge 1 : 0.7222, Rouge 2 : 0.5882, Rouge L : 0.7222 |
| 4 | Bagaimana Pengurusan Surat Keterangan Aktif Kuliah? | Pengurusan Surat Keterangan Aktif Kuliah silahkan ke bagian Registrasi. | Silahkan ke bagian Registrasi. | Rouge 1 : 0.6153, Rouge 2 : 0.5454, Rouge L : 0.6153 |
| 5 | Saya Mahasiswa Prodi AI dan saya lupa Password atau kata sandi KRS Online kemana saya harus melapor? | Mahasiswa lupa Password atau kata sandi KRS Online, terkait hal ini Mahasiswa dapat membuat permohonan melalui laman website di : ult.usk.ac.id atau langsung melalui link google form : <https://forms.gle/Xq6ThdgQNhWYhbat8> | Apabila mahasiswa lupa password atau kata sandi KRS, dapat melakukan permohonan ke Unit Layanan Terpadu (ULT) di Biro USK. Mereka akan membantu mengatur ulang penggunaan akun dan menyediakan informasi tentang bagaimana cara memperbarui password. | Rouge 1 : 0.3582, Rouge 2 : 0.1846, Rouge L : 0.3283 |

Tabel 4.9. Nilai skor ROUGE dengan Metode RAG

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metode | Jumlah Pertanyaan | Skor ROUGE | | |
| R-1 | R-2 | R-L |
| RAG | 15/56 | >0.5 | >0.5 | >0.5 |

## Kategori skor ROUGE

Skor Rouge yang baik bervariasi berdasarkan tugas pada ringkasan dan metrik. Skor Rouge-1 dengan kategori sangat baik dengan nilai sekitar 0,5, dan skor di atas 0,5 dianggap baik dan 0,4 hingga 0,5 sedang. Untuk Rouge-2, skor di atas 0,4 masuk ke dalam kategori baik, dan 0,2 hingga 0,4 masuk ke dalam kategori sedang.

Skor Rouge-L dengan kategori bagus mendapatkan nilai sekitar 0,4 dan kategori rendah berkisar pada nilai 0,3 hingga 0,4. Meskipun skor Rouge berguna, skor tersebut tidak memperhitungkan kualitas semantik atau sintaksis dan harus dilengkapi dengan metrik lain dan evaluasi manusia untuk penilaian yang lengkap (Walker II, 2024).

Tabel 4.10. Tabel Kategori Nilai Metrik Rouge (Walker II, 2024)

|  |  |  |  |
| --- | --- | --- | --- |
| Metrik Rouge | Baik sekali | Baik | Sedang |
| Rouge-1 | 0.5+ | >0.5 | 0.4-0.5 |
| Rouge-2 | - | >0.4 | 0.2-0.4 |
| Rouge-L | - | ~0.4 | 0.3-0.4 |

## Menghitung Evaluasi Sumber Daya

Pada penelitian ini peneliti melakukan uji waktu pada model Mistral 7B, dengan pengujian *fine-tuning* dan inferensi. Peneliti menggunakan satu kartu grafis NVIDIA Tesla T4 GPU yang tersedia di Google Colab untuk inferensi dan juga untuk *fine-tuning*. Berdasarkan uji waktu *fine-tuning* dengan menggunakan 20 dataset, seperti yang ditunjukkan pada Tabel 4.11, Peneliti menyimpulkan bahwa pendekatan *fine-tuning* dengan library *unsloth* efektif dan efisien, serta membutuhkan waktu yang minimal. Pengujian waktu inferensi menunjukkan bahwa USK Mistral 7B tidak memerlukan daya komputasi yang berlebihan selama proses inferensi, sehingga menjadikannya sistem hemat energi dan berefisiensi untuk merespons pertanyaan dari pengguna hanya saja pengujian menggunakan metode RAG membutuhkan waktu 5-6 menit untuk merespons pertanyaan pengguna. Harapannya model ini akan menjadi sistem yang efisien dalam menangani informasi terkait administrasi akademik USK di masa mendatang.

Tabel 4.11. Hitungan waktu pada model saat *fine-tuning* dan menjalankan RAG

|  |  |  |
| --- | --- | --- |
| Model | Waktu *Fine-tuning* (jam) | Waktu RAG (menit) |
| USK Mistral 7B | 2 | 5-6 |

## Analisis Hasil

### Masalah Data Pelatihan

Faktor penting yang berkontribusi terhadap halusinasi LLM adalah sifat data pelatihan. LLM, seperti pada model Mistral 7B dan Llama2, menjalani pelatihan ekstensif tanpa pengawasan dengan kumpulan data yang besar dan beragam dari berbagai asal (Deval, 2023a). Memverifikasi kebenaran faktual data ini merupakan sebuah tantangan. Saat model ini belajar menghasilkan teks, model tersebut juga dapat menemukan dan mereplikasi ketidakakuratan faktual dalam data pelatihan. Hal ini mengarah pada skenario dimana model tidak dapat membedakan antara kebenaran dan fiksi dan dapat menghasilkan keluaran yang menyimpang dari fakta atau penalaran logis. LLM yang dilatih berdasarkan kumpulan data yang bersumber dari internet dapat berisi informasi yang bias atau salah. Informasi yang salah ini dapat menyebar ke keluaran model, karena model tidak dapat membedakan antara data yang akurat dan tidak akurat.

### Mengurangi Halusinasi

Upaya untuk mengurangi halusinasi sangat penting untuk menjaga kredibilitas dan fungsionalitas LLM. Metode utama untuk mengidentifikasi dan mengurangi kesalahan ini melibatkan kombinasi metrik canggih dan evaluasi kritis manusia. Ini termasuk:

1. Metrik kualitas linguistik seperti ROUGE dan BLEU.
2. Metrik validitas konten, yaitu berbasis IE, berbasis QA, dan berbasis NLI (Minaee et al., 2024).
3. *FactScore* untuk memeriksa keakuratan fakta individu (Deval, 2023b).

### Metode untuk mendeteksi halusinasi pada LLM

Metode inovatif seperti *SelfCheckGPT* mendeteksi halusinasi dengan menilai konsistensi beberapa jawaban yang dihasilkan untuk pertanyaan yang sama. Selain itu, teknik seperti *chain-of-thought prompting* dan RAG terus dieksplorasi untuk memperkuat kemampuan model dalam memberikan informasi yang tepat dan relevan (Deval, 2023b).

### Pengaruh GPU dalam Implementasi LLM

Dalam menjalankan LLM, GPU memegang peranan penting. GPU khusus dengan VRAM tinggi dapat mempercepat komputasi yang dibutuhkan oleh model secara signifikan. Pada penelitian ini GPU yang digunakan yaitu “NVIDIA Tesla T4 GPU” yang tersedia di Google Colab secara gratis, hasil pengujian dengan GPU ini menggunakan metode RAG membutuhkan waktu yang cukup lama untuk menghasilkan respons dari pertanyaan yang diajukan.

# BAB V KESIMPULAN DAN SARAN

## Kesimpulan

Chatbot untuk mendapatkan informasi akademik sudah berhasil dibangun dengan akurasi skor Rouge>0.5 pada Rouge-1, Rouge-2, dan Rouge-L. Skor Rouge ini didapat dari pengujian menggunakan 56 pertanyaan terkait akademik di USK, Sebanyak 15 pertanyaan mendapatkan skor yang baik yaitu dengan skor >0.5.

Perbandingan kinerja antara *fine-tuning* dan RAG menunjukan bahwa skor Rouge-1, Rouge-2, dan Rouge-L pada *fine-tuning* mendapatkan nilai skor yang baik yaitu >0.5. Skor ini didapat dengan pengujian pada model dengan memberikan pertanyaan yang berbeda tetapi memiliki makna yang sama. Kinerja menggunakan metode RAG menghasilkan skor pada Rouge-1, Rouge-2, dan Rouge-L dengan nilai >0.5 diuji pada 15 dari 56 pertanyaan yang diajukan.

File informasi akademik dalam bentuk pernyatan berhasil dibangun dengan jumlah pernyataan sebanyak 231 data informasi akademik di USK. Dokumen ini digunakan pada chatbot dengan metode RAG. Hasilnya memberi peningkatan waktu pada model untuk menghasilkan respon terkait topik di bidang Akademik dan meningkatkan akurasi pada model.

## Saran

Disarankan untuk penelitian selanjutnya selain menambah dataset agar model menjadi lebih handal. Penelitian juga dapat dilakukan dengan pengujian pada model-model lain yang lebih baru dengan fitur-fitur yang lebih canggih. Dengan menjelajahi hal tersebut maka didapatkan hasil perbandingan pada model-model lain sehingga ditemukan model yang terbaik untuk domain pelayanan Administrasi Akademik.

# DAFTAR KEPUSTAKAAN

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 4895–4901. https://doi.org/10.18653/v1/2023.emnlp-main.298

Amanat, M. U. (2024). *LLM evaluation with Rouge*. Medium.Com. https://medium.com/@MUmarAmanat/llm-evaluation-with-rouge-0ebf6cf2aed4

Andryanto, S. D. (2021). *Apa itu Chatbot? Begini Cara Kerja Asisten Digital ini*. Tekno.Tempo.Co. https://tekno.tempo.co/read/1536593/apa-itu-chatbot-begini-cara-kerja-asisten-digital-ini

Baker, H. P., Dwyer, E., Kalidoss, S., Hynes, K., Wolf, J., & Strelzow, J. A. (2024). ChatGPT’s Ability to Assist with Clinical Documentation: A Randomized Controlled Trial. *Journal of the American Academy of Orthopaedic Surgeons*, *32*(3), 123–129. https://doi.org/10.5435/JAAOS-D-23-00474

Barandoni, S., Chiarello, F., Cascone, L., & Puccio, S. (2024). *Automating Customer Needs Analysis : A Comparative Study of Large Language Models in the Travel Industry*.

Benveniste, D. (2023). *Augmenting LLMs: Fine-Tuning or RAG?* Newsletter.Theaiedge.Io. https://newsletter.theaiedge.io/p/augmenting-llms-fine-tuning-or-rag

Bhatti, A., Parmar, S., & Lee, S. (2023). *SM70: A Large Language Model for Medical Devices*. *1*, 1–5.

data.usk.ac.id. (2024). *Data Mahasiswa Daftar Ulang*. Data.Usk.Ac.Id. https://data.usk.ac.id/mahasiswa-daftar

Deval, S. (2023a). *The Beginner’s Guide to Hallucinations in Large Language Models*. https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models#:~:text=Mitigating hallucinations in LLMs involves,implement various product design strategies

Deval, S. (2023b). *The Beginner’s Guide to Hallucinations in Large Language Models*.

diskominfo.kedirikab.go.id. (2023). *Apa itu bahasa Python?* Diskominfo.Kedirikab.Go.Id. https://diskominfo.kedirikab.go.id/baca/apa-itu-bahasa-python

docs.llamaindex.ai. (2023). *High-Level Concepts*. Docs.Llamaindex.Ai. https://docs.llamaindex.ai/en/v0.10.17/getting\_started/concepts.html

Fatyanosa, T. (2020). *Fine-Tuning Pre-Trained Transformer-based Language Model*. Medium.Com. https://fatyanosa.medium.com/fine-tuning-pre-trained-transformer-based-language-model-c542af0e7fc1

Grootendorst, M. (2023). *Which Quantization Method is Right for You? (GPTQ vs. GGUF vs. AWQ)*. Maartengrootendorst.Com. https://www.maartengrootendorst.com/blog/quantization/

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-Rank Adaptation of Large Language Models. *ICLR 2022 - 10th International Conference on Learning Representations*, 1–26.

Huang, Q., Tao, M., Zhang, C., & An, Z. (2023). *Lawyer LLaMA: Enhancing LLMs with Legal Knowledge*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. Le, Lavril, T., Wang, T., Lacroix, T., & Sayed, W. El. (2023). *Mistral 7B*. 1–9. http://arxiv.org/abs/2310.06825

Jonatan, V., & Igor, and A.-A. (2023). Creation of a Chatbot Based on Natural Language Processing for Whatsapp. *Journal of Database Management*, *3*(4), 39–53. https://doi.org/10.14810/elelij.2023.12402

Khan, A. A. (2023). *“Crafting Conversational Magic: Building an AI Language Model(LLM) App with Langchain?and Gradio.”* Medium.Com. https://medium.com/@abirkhan4u/crafting-conversational-magic-building-an-ai-language-model-llm-app-with-langchain-and-gradio-496429567512

Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew*. 74–81. https://doi.org/10.1253/jcj.34.1213

Loukas, L., Stogiannidis, I., Malakasiotis, P., & Vassos, S. (2023). Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance. *FinNLP-Muffin 2023 - Joint Workshop of the 5th Financial Technology and Natural Language Processing and 2nd Multimodal AI For Financial Forecasting, in Conjunction with IJCAI 2023 - Proceedings*, 74–80.

Martin, M. (2023). *Large Language Models (LLMs): Definition, Examples, and Benefits*. Smith.Ai. https://smith.ai/blog/llm-ai

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey*. http://arxiv.org/abs/2402.06196

Minsky, M. (1989). *Steps Toward Artificial Intelligence*. Web.Media.Mit.Edu. https://web.media.mit.edu/~minsky/papers/steps.html

Mohamadi, S., Mujtaba, G., Le, N., Doretto, G., & Adjeroh, D. A. (2023). *ChatGPT in the Age of Generative AI and Large Language Models: A Concise Survey*. 1–60. http://arxiv.org/abs/2307.04251

Neves, M. C. (2023). *What are Quantized LLMs?* Https://Www.Tensorops.Ai/. https://www.tensorops.ai/post/what-are-quantized-llms

Rakotoson, Loïc, Massip, Sylvain, A. A. Laleye, F. (2024). *Retrieval Augmented Generation Architecture*. Www.Researchgate.Net. https://www.researchgate.net/figure/Retrieval-Augmented-Generation-Architecture\_fig1\_378364457

Revou.co. (2024). *Apa itu Natural Language Processing (NLP)*. Revou.Co. https://revou.co/kosakata/natural-language-processing

Risyad, S. A. (2023). *Data Set: Pengertian, Jenis, dan Contohnya*. Dibimbing.Id. https://dibimbing.id/blog/detail/pengertian-data-sheet-jenis-dan-contoh

Russel, Stuart J; Norvig, P. (2010). *Artificial intelligence :a modern approach*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. http://arxiv.org/abs/2307.09288

Trozze, A., Davies, T., & Kleinberg, B. (2023). *Large Language Models in Cryptocurrency Securities Cases: Can ChatGPT Replace Lawyers?* 1–49. http://arxiv.org/abs/2308.06032

ult.usk.ac.id. (2024). *Unit Layanan Terpadu USK*. Ult.Usk.Ac.Id. http://ult.usk.ac.id/

V., K. (2024). *Fine-Tuning Large Language Models with Unsloth*. Medium.Com. https://medium.com/@kushalvala/fine-tuning-large-language-models-with-unsloth-380216a76108

Walker II, S. M. (2024). *What is the ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)?* Klu.Ai. https://klu.ai/glossary/rouge-score

Wicaksono, K. T. (2023). *Memahami Retrieval Augmented Generation (RAG) Melalui Rencana Pembangunan LaporMacet.com*. Medium.Com. https://kukuhtw.medium.com/memahami-retrieval-augmented-generation-rag-melalui-rencana-pembangunan-lapormacet-com-5990a76edda7

www.gradio.app. (2024). *Gradio*. Www.Gradio.App. https://www.gradio.app/

www.wiz.ai. (2023). *Menelaah : Bagaimana LLM (Large Language Models) bekerja*. Www.Wiz.Ai.

Zhao, H., Ling, Q., Pan, Y., Zhong, T., Hu, J.-Y., Yao, J., Xiao, F., Xiao, Z., Zhang, Y., Xu, S.-H., Wu, S.-N., Kang, M., Wu, Z., Liu, Z., Jiang, X., Liu, T., & Shao, Y. (2023). Ophtha-LLaMA2: A Large Language Model for Ophthalmology. *ArXiv [Preprint]*, 1–19. http://arxiv.org/abs/2312.04906

LAMPIRAN

Lampiran 1. Sampel dataset Q&A di USK

|  |  |  |  |
| --- | --- | --- | --- |
| No | Label | Pertanyaan | Jawaban |
| 1 | Umum | Bagaimana cara mendapatkan Sertifikat Akreditasi Universitas? | Untuk Sertifikat Akreditasi silakan download disini ya https://lpm.usk.ac.id/, untuk Sertifikat Akreditasi yang terbaru tidak perlu dilegalisir karena sudah ada QR Code. |
| 2 | Umum | Bagaimana jika lupa password untuk pengisian KRS Online? | Terkait dengan riset kata sandi KRS Online, Anda bisa membuat permohonan melalui laman website Kami di: ult.usk.ac.id atau langsung melalui link google form: https://forms.gle/Xq6ThdgQNhWYhbat8. Jika ada kendala, Anda bisa menghubungi Kami via email di ult@usk.ac.id (khusus untuk riset password KRS, riset password email domain USK, perbaikan biodata KRS, dan perbaikan data PDDIKTI). |
| 3 | UKT | Bagaimana cara membayar UKT untuk jalur SMMPTN Barat, dan UKT nya apakah di bayar sekalian dengan IPI nya? | Silahkan langsung bayar ke Bank, dan masalah IPI tidak boleh di cicil kecuali hanya Fakultas Kedokteran saja. |
| 4 | UKT | Kapan jadwal sanggahan UKT bagi Mahasiswa? | Untuk data UKT yang sudah dibuat tidak bisa diubah kembali dan data UKT itu tidak menjadi masalah yang penting data KRS benar semua dan mengenai banding UKT nanti akan dibuka oleh bagian Keuangan, silakan pantau infonya di web usk.ac.id. |
| 5 | Akademik | Bagaimana cara membuat Surat Keterangan Aktif Kuliah? | Terkait masalah Surat Keterangan Aktif Kuliah silahkan mengurus ke bagian Registrasi Akademik. |
| 6 | Umum | Dimana informasi mengenai Jadwal Daftar Ulang Mahasiswa Baru jalur SNBP dan SNBT? | Link Jadwal Pendaftaran Ulang Calon Mahasiswa Baru jalur SNBP dan SNBT https://usk.ac.id/tahapan-pendaftaran-ulang-jadwal-dan-tata-cara-calon-mahasiswa-jalur-snbt-2023/. Mohon dibaca dengan seksama agar tidak ada informasi yang tertinggal. |
| 7 | UKT | Kenapa di dalam pengisian UKTB dinyatakan belum lolos validasi? | Kolom validasi ini, harus menunggu sampai Status Buta Warna berubah menjadi Warna Hijau. |
| 8 | Umum | Apakah boleh jika menggunakan sertifikat TOEFL dari penyelenggara tes selain USK? | Sertifikat yang diakui oleh USK adalah ITP dan akan divalidasi dengan pihak pusat ITP. Untuk prosesnya dapat memperlihatkan bukti fisik dan akan divalidasi oleh Lab Bahasa. |
| 9 | Umum | Bagaimana cara mendaftar Bimbingan/Kelas TOEFL? | Silahkan CP ke WA Only UPT. Bahasa secara online di 081375386018, Pelatihan akan dibuka jika sudah memenuhi kuota kelas, Pendaftaran TOEFL wajib menyertakan Surat Pengantar yang dikeluarkan oleh Prodi masing-masing. |
| 10 | Umum | Apakah mahasiswa bisa pindah ke Fakultas Lain? | Mahasiswa tidak bisa pindah Fakultas. Yang boleh pindah hanya program studi saja. |
| 11 | Akademik | Bagaimana cara mengurus Cuti Perkuliahan? | Pengurusan Cuti Semester dilakukan di bagian Akademik Fakultas masing-masing. |
| 12 | Akademik | Bagaimana cara Cetak KTM? | KTM dapat dicetak di laman  berkas-akademik.usk.ac.id. |
| 13 | Umum | Bagaimana Jika Alamat KK dan KTP berbeda dengan alamat tempat tinggal sekarang? | Tetap gunakan alamat yang sekarang yang menjadi tempat tinggal Anda. Jika nanti ada permasalahan, maka tim verifikasi sendiri yang akan menghubungi untuk meminta penjelasan mengenai perbedaan alamat sekarang dan di KTP & KK. |
| 14 | UKT | Kenapa UKTB bisa sangat mahal? | Sistem membaca apa yang Anda isi. Jika merasa keberatan, silahkan ikut Peninjauan UKT. Informasinya tersedia di laman: usk.ac.id. Untuk UKT semester ini tetap harus dibayarkan agar tidak dianggap mengundurkan diri. |
| 15 | UKT | Bagaimana jika terlambat melakukan Pembayaran Uang Kuliah? | Jika telat melakukan pembayaran UKT, Silahkan ke bagian Akademik untuk meminta dibuatkan surat dari WD 1 yang ditujukan ke WR 1 dimana isi suratnya yaitu minta dibukakan kembali pembayaran UKT. |
| 16 | Umum | Bagaimana jika Deposit Uang Toga belum dikembalikan? | Setelah mengisi data pada link yang telah diberikan, mohon bersabar karena pasti akan di transfer oleh pihak Keuangan USK. Saat ini memang sedang dalam tahap proses pen-transferan. |
| 17 | Umum | Bagaimana jika kehilangan bukti SPP 6 semester terakhir? | Silahkan membuat permohonan melalui laman website di ult.usk.ac.id. Jika ingin dibuatkan slip SPP pengganti hanya 1 atau 2 semester saja, maka tidak perlu bayar. Lebih dari itu maka harus membayar sebesar Rp.2.000/lembar atau per semester. |
| 18 | Umum | Bagaimana cara membuat Slip SPP Pengganti? | Terkait dengan pembuatan Slip SPP Pengganti dapat membayar sebesar Rp 18.000,-. Berikut Kami lampirkan nomor rekening: Nama Bank : Bank Syariah Indonesia No. Rekening : 914-471-111-122-001-6 Bukti pembayarannya, upload di laman website Kami : ult.unsyiah.ac.id Pilih Layanan kemudian pilih Slip SPP Pengganti untuk jadwal pengambilan Slip SPP silahkan hubungi Unit Layanan Terpadu. |
| 19 | Umum | Bagaimana cara membuat Surat Rekomendasi Kampus untuk Pengurusan Beasiswa? | Pembuatan Surat Rekomendasi Kampus dapat dilakukan di bagian Kemahasiswaan pada Fakultas masing-masing. |
| 20 | Umum | Bagaimana jika KRS belum difinalisasi oleh Dosen Wali? | Dapat melapor hal ini pada Program Studi masing-masing. |

Lampiran 2. Sampel dataset untuk metode RAG

|  |  |
| --- | --- |
| No | Data Informasi USK |
| 1 | KTM yang belum dapat dicetak dapat dilihat melalui laman website <https://berkas-akademik.usk.ac.id/> untuk mengecek kesesuain data atau melapor ke Bagian Unit Layanan Terpadu (ULT) di Biro USK. |
| 2 | Keterlambatan dalam pembayaran Uang Kuliah Atau UKT dapat meminta ke Fakultas untuk dibuatkan surat dari WD 1 yang ditujukan ke WR 1 dimana isi surat tersebut berisi permohonan untuk meminta dibukakan kembali pembayaran UKT kemudian surat tersebut diberikan ke bagian Tata Usaha di Biro Rektor USK. |
| 3 | Pengajuan Surat Rekomendasi Kampus dapat dilakukan di bagian Kemahasiswaan. |
| 4 | KRS yang belum terfinalisasi dapat melapor melalui Admin pada Program Studi masing masing. |
| 5 | Link informasi mengenai Biaya Kuliah atau UKTB di USK Mahasiswa dapat mengunjungi pada link berikut ini  <https://usk.ac.id/download/uktb-mahasiswa-s1-jalur-snbp-snbt-dan-mandiri-usk-2023/>. |
| 6 | Pemesanan Almamater USK untuk Mahasiswa Baru informasinya silahkan untuk mengunjungi laman <https://pakarmaru.usk.ac.id/> atau konsultasi ke ULT. |
| 7 | Pengajuan permohonan *Letter of Admission/Acceptance* (LoA) dapat mengirimkan berkas ke email “[persuratan@usk.ac.id](mailto:persuratan@usk.ac.id)”. |
| 8 | Data nama/tanggal lahir/alamat orang tua dapat diperbaiki pada saat pengisian Biodata Online pada KRS Online. |
| 9 | Jika merasa keberatan dengan besaran UKTB maka silahkan untuk melakukan verifikasi data UKTB pada jadwal yang telah ditentukan (lihat jadwal dan tempat di website usk.ac.id) dengan membawa bukti/dokumen pendukung. |
| 10 | Pembayaran Uang Kuliah dapat dilakukan satu hari setelah melakukan pengisian UKTB dengan membawa nomor pendaftaran SNMPTN jika dilakukan via teller bank mitra USK. |