JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

IMPLEMENTASI LOW-RANK ADAPTATION OF LARGE LANGUANGE MODEL (LORA) UNTUK EFFISIENSI LARGE LANGUAGE MODEL

Anton Mahendra*[1)], Styawati[2)]
1. Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia
2. Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia

Article Info
Kata Kunci: Effisiensi, LoRA, Model bahasa, LlaMA2-7B, *Streamlit*

Keywords: *Efficiency, LoRA, Languange Model, LlaMA2-7B, Streamlit*

Article history:
Received 13 September 2024
Revised 10 Oktober 2024
Accepted 4 November 2024
Available online 4 December 2024

DOI :
https://doi.org/10.29100/jipi.v9i4.5519

* Corresponding author.
Corresponding Author
E-mail address:
anton_mahendra@teknokrat.ac.id

ABSTRAK
Model transformator seperti LlaMA 2-7B sangat kuat untuk memproses berbagai tugas bahasa alami, namun memiliki kekuatan pemrosesan yang signifikan dan keterbatasan memori yang membuatnya sulit untuk diimplementasikan. Tantangan terbesarnya terletak pada konsumsi sumber daya penyimpanan yang besar dan kebutuhan daya komputasi dalam jumlah besar. Untuk mengatasi permasalahan tersebut, dikembangkan solusi berupa implementasi LoRA (*Low Rank Adapter*). LoRA, khususnya di LlaMA 2-7B, menggunakan pendekatan adaptif dalam mengompresi model *Transformer* menggunakan adaptor berdaya rendah. Penerapan LoRA pada model ini mengurangi jumlah operasi *floating-point*, sehingga mempercepat proses pelatihan dan inferensi. Secara signifikan mengurangi konsumsi daya dan penggunaan memori. Tujuan utama penerapan LoRA di LlaMA 2-7B adalah untuk mengoptimalkan efisiensi model, dengan fokus pada pengurangan operasi floating-point dan meningkatkan penggunaan memori GPU.

ABSTRACT
Transformer models such as LlaMA 2-7B are powerful for processing various natural language tasks, but have significant processing power and memory limitations that make them difficult to implement. The biggest challenge lies in the consumption of large storage resources and the need for large amounts of computing power. To overcome these problems, a solution was developed in the form of LoRA (Low Rank Adapter) implementation. LoRA, especially in LlaMA 2-7B, uses an adaptive approach in compressing the Transformer model using low-rank adapters. The implementation of LoRA in this model reduces the number of floating-point operations, thus speeding up the training and inference process. It significantly reduces power consumption and memory usage. The main objective of applying LoRA in LlaMA 2-7B is to optimise the efficiency of the model, with a focus on reducing floating-point operations and improving GPU memory usage.

I. PENDAHULUAN

PERKEMBANGAN model bahasa besar (LLM) telah mengalami kemajuan signifikan dalam beberapa tahun terakhir dan telah menjadi pilar transformasi sosial *modern*. Pada tahun 2023, Indonesia mencapai peringkat ketiga dalam adopsi kecerdasan buatan (AI), dengan kontribusi pengguna internet Indonesia mencapai 5,60% dari total 1,4 milliar akses AI dalam lalu lintas dunia. Model bahasa besar merupakan model *deep learning* yang sangat besar dan telah dilatih sebelumnya dengan sejumlah besar data [1]. Model bahasa besar, seperti *Google AI* versi 2022 dengan PaLM2 dengan 540 miliar parameter, kini dapat menyelesaikan tugas-tugas kompleks seperti menulis kode dan menjawab pertanyaan yang diberikan [2]. Sebagai model bahasa besar terkemuka, PaLM2 mengubah paradigma tradisional pra-pemrosesan bahasa berskala besar dan membuka potensi baru yang sebelumnya tidak terbayangkan.

Pemrosesan model skala besar juga dapat mempelajari pola dan hubungan kompleks dalam bahasa, sehingga memungkinkan untuk melakukan berbagai hal yang sebelumnya memerlukan intervensi manusia, seperti menerjemahkan bahasa, menyusun kalimat, dan menjawab pertanyaan [3]. Pemrosesan model bahasa besar telah digunakan untuk merevolusi banyak bidang, termasuk generasi teks. Dalam penelitian ini, peneliti menggunakan pemrosesan bahasa skala besar, khususnya model LlaMA 2-7B, untuk menganalisis data survei teks, membuat teks kreatif, dan menjawab pertanyaan tentang data kesehatan [3].

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

Pemrosesan bahasa skala besar diimplementasikan menggunakan teknik Adaptasi Tingkat Rendah (LoRA) yang meningkatkan efisiensi waktu komputasi. LoRA menggunakan algoritma tingkat rendah dari berbagai sumber untuk mengatasi kelemahan waktu komputasi yang relatif lama dalam pemrosesan model skala besar, yang terutama bergantung pada kompleksitas tugas yang sedang diproses [4]. LoRA membantu model bahasa mengurangi waktu komputasi yang diperlukan untuk menjawab pertanyaan dengan memberikan skor lebih tinggi pada kata-kata yang dianggap lebih penting dalam sebuah pertanyaan [5].

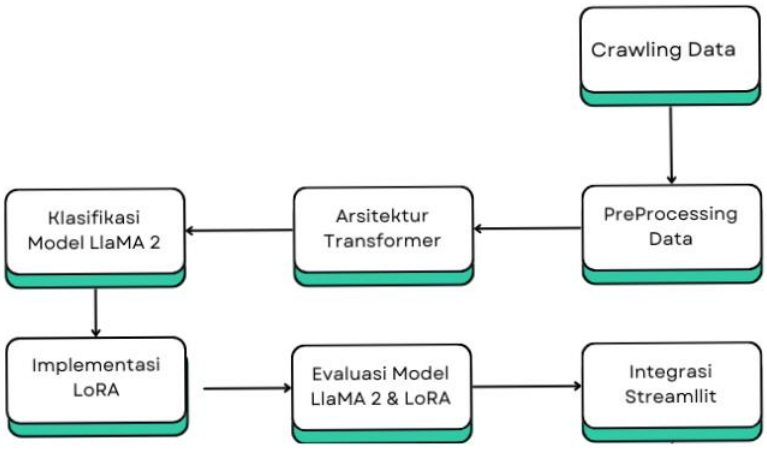
Manfaat LoRA mencakup aspek yang meningkatkan efisiensi dan fleksibilitas penggunaan [4]. Pertama, model terlatih dapat dipartisi dan digunakan untuk membuat modul LoRA kecil yang mendukung berbagai tugas [4]. Hal ini memungkinkan peneliti untuk membekukan model bersama-sama dan mengalihkan tugas secara efisien, sehingga mengurangi kebutuhan memori dan *overhead* peralihan tugas [6]. Kedua, dengan menggunakan optimasi adaptif, LoRA meningkatkan efisiensi pelatihan dan mengurangi kemacetan perangkat keras sebanyak 3 kali lipat [6]. Menghilangkan penghitungan gradien dan mempertahankan status pengoptimal untuk sebagian besar parameter akan mempercepat proses pelatihan dengan memungkinkan peneliti fokus pada pengoptimalan matriks yang lebih kecil dan berperingkat lebih rendah [7].

Dalam upaya menyempurnakan model LlaMA 2-7B, peneliti memilih varian LlaMA 2-7B ini karena memiliki sejumlah keunggulan yang signifikan. Kemampuannya untuk menghasilkan energi yang tidak terbatas, efisiensi dalam penggunaan energi, adaptasi lingkungan, dan fitur keamanan yang tinggi menjadikannya pilihan yang tepat. Dalam perbandingan dengan model bahasa besar lainnya, LlaMA2 -7B dalam generasi teks, skala yang sangat besar, kinerja. Relevansinya dengan konteks penelitian terletak pada kemampuannya untuk dioptimalkan dengan metode adaptasi efisien, seperti adaptor LlaMA 2-7B, yang membantu mengoptimalkan sumber daya komputasi. Namun, kelemahan dari model ini yang dimiliki oleh LlaMA2-7B ialah kompleksitas yang tinggi dan kebutuhan akan memori GPU yang besar. Meskipun demikian, dengan pengembangan metode efisien, kelemahan ini dapat diatasi. Oleh karena itu, sesuai dengan fokus penelitian pada aplikasi LLM yang dioptimalkan dengan metode adaptasi efisien, menunjukkan potensi besar LlaMA 2-7B dalam memajukan penelitian ini dan memperluas aplikasi model bahasa dalam berbagai bidang.

Penelitian ini bertujuan menyempurnakan model LlaMA2 -7B dengan metode adaptasi efisiensi dan LoRA untuk mengoptimalkan sumber daya komputasi. Tujuan utama adalah meningkatkan kinerja aplikasi pemrosesan bahasa alami di Indonesia. Diharapkan penelitian ini akan mendukung pengembangan teknologi bahasa alami di di Indonesia dan memperluas akses masyarakat terhadap aplikasi yang menggunakan teknologi tersebut.

II. METODELOGI

Metode penelitian merujuk pada serangkaian kegiatan yang dilaksanakan dengan metode yang teratur guna mencapai tujuan penelitian. Metode penelitian terlihat pada Gambar 1.



Gambar. 1. Metode Penelitian

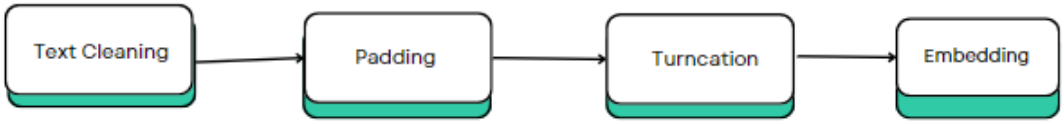
A. Pengumpulan Data

pengumpulan data dilakukan dengan menggunakan metode *web scrapping* pada salah satu situs *platform* yaitu *Apify* (*https://apify.com/)* dan juga pustaka dari *python* yaitu *beautifulsoup* dengan menyalin URL artikel kesehatan. Data yang dipakai dalam proses implementasi LoRA untuk model bahasa skala besar adalah artikel terkait dengan pengetahuan umum berupa sejarah, Pendidikan, Kesehatan, serta teknologi, penyebabnya, serta deskripsi terkait artikel tersebut. Hasil tersebut kemudian disimpan dalam bentuk file .csv (*comma separated values*).

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

B. Preprocessing

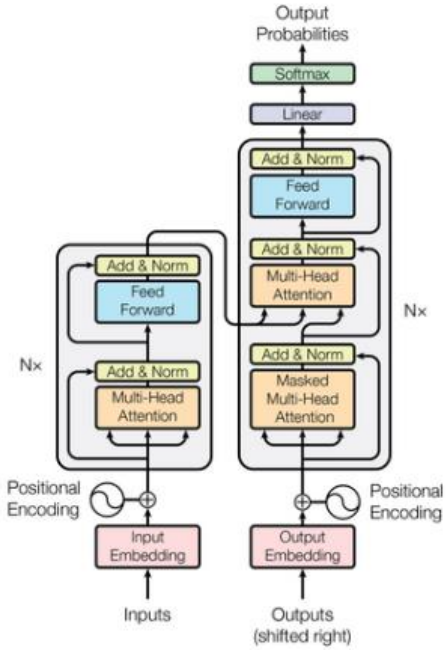
Pada tahap Preprocessing data mentah yang telah dipersiapkan untuk digunakan melalui proses *preprocessing.* Pengolahan data dengan *preprocessing* melibatkan berbagai tahapan, seperti *text cleaning*, *padding*, *turncation*, dan *embedding* [8]. Tahapan *preprocessing* data ditunjukkan pada Gambar 2.



Gambar. 2. Metode Penelitian

C. Arsitektur *Transformer*

Arsitektur *Transformer*, khususnya varian LoRA (*Low Rank Adaptation of Large Language Model*), merupakan perkembangan menarik dalam dunia pemrosesan bahasa alami. Dibandingkan pendekatan sebelumnya, LoRA lebih fokus pada bagian *encoder* pada arsitektur *Transformer,* dimana mekanisme *self-attention* menjadi inti utamanya [6]. Dengan memanfaatkan perhatian diri, LoRA dapat memberikan bobot yang sesuai untuk setiap kata dalam sebuah kalimat, sehingga model dapat lebih memahami konteks keseluruhan [9]. Selain itu, parameter arsitektur yang dapat dilatih ini memungkinkan model menangkap pola dan fitur kompleks dalam data, sehingga menghasilkan representasi linguistik yang lebih dalam dan akurat [10]. Menggabungkan mekanisme perhatian tingkat lanjut dan kemampuan eksekusi paralel, LoRA memberikan dukungan kuat untuk pemahaman bahasa dan pemrosesan data yang efisien [6]. Hal ini membuat arsitektur LoRA sangat menjanjikan untuk meningkatkan berbagai aspek pemrosesan bahasa alami dan membuka pintu bagi kemajuan lebih lanjut di masa depan.



Gambar. 3. Arsitektur *Transformer*

D. Klasifikasi Model LlaMA 2-7B 7B

Setelah membuat arsitektur selanjutnya membuat model untuk mengklasifikasikan teks untuk generasi teks. LlaMA 2 merupakan penerus dari model Llama, yang merupakan model bahasa besar yang dilatih oleh META AI. Model ini membantu memahami dan merespons masukan manusia serta menghasilkan teks yang mirip dengan bahasa manusia. Kemampuan klasifikasi untuk pembuatan teks menggunakan model bahasa seperti Llama2 bertujuan untuk mengontrol pembuatan teks untuk menghasilkan keluaran yang sesuai dengan kelas atau label tertentu [11]. Langkah pertama melibatkan persiapan data. Di sini, kumpulan data terkait dikumpulkan dan diklasifikasikan menurut kelas atau label yang diinginkan [12]. Setelah itu model dilatih dengan *fine-tune*. Skala model ini sebesar 7 miliar parameter, yang kemudian disebut LlaMA2-7B dioptimalkan untuk kasus penggunaan dialog. Selanjutnya, model tersebut kemudian dapat digunakan untuk menghasilkan teks yang sesuai dengan kelas atau label yang diinginkan. Oleh karena itu, fungsi klasifikasi berperan penting dalam mengontrol

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

pembuatan teks untuk menghasilkan keluaran yang memenuhi tujuan dan kebutuhan tertentu [11]. Dengan dipilihnya adaptor LlaMA2 7B sebagai bagian dari seri model bahasa besar, mengadopsi arsitektur transformer yang dirancang untuk generasi teks secara *auro-regressive*. Arsitektur ini telah melalui *fine-tuning* dan menggunakan teknik *Supervised-Fine-Tuning* (SFT) dan *Reinforcement Learning with Human Feedback* (RLHF), dengan tujuan untuk menyesuaikan model dengan prefensi manusia terkait kebermanfaatan dalam menghasilkan teks [13].

E. Implementasi LoRA

Saat mengimplementasikan LoRA, peneliti menentukan rasio kompresi yang optimal untuk model LlaMA2 7B. Langkah pertama adalah memilih model dasar yang kompleks seperti LlaMA2 7B dan menganalisis struktur dan parameternya dengan skala yang besar dan efisiensi dalam pemahaman bahasa karena relevansinya untuk pengembangan model yang berkinerja tinggi dengan sumber daya terbatas[4]. Kemudian menerapkan teknik peringkat rendah untuk mengurangi dimensi parameter model, mencapai efisiensi sumber daya tanpa mengurangi kualitas pemahaman bahasa [7]. Rasio kompresi ditentukan dan akurasi serta kecepatan inferensi model terkompresi diuji setelah implementasi. Disini peneliti memilih menggunakan rasio kompresi 4x, yang artinya pengurangan dimensi parameter sebanyak 4 kali lipat. Ini akan menguranngi jumlah parameter yang harus diperbarui selama pelatihan, sehingga menghemat memori dan waktu komputasi.[14]. Implementasi LoRA di LlaMA2 7B bertujuan untuk menciptakan model bahasa efisien yang relevan dengan aplikasi dengan sumber daya terbatas[4]. Selanjutnya, evaluasi kinerja model sesudah penerapan LoRA menjadi sangat penting. Hal ini memberikan pemahaman yang lebih baik tentang efektivitas teknik ini. Dengan implementasi LoRA, harapannya model LlaMA2 7B dapat mencapai efisiensi sumber daya dan berkinerja tinggi dalam aplikasi terbatas.

F. Evaluasi

Selama fase evaluasi model LlaMA 2-7B dan penerapan teknik LoRA, langkah-langkah penting diambil untuk mengevaluasi kinerja dan efisiensi model. Evaluasi dimulai dengan mengukur akurasi model LlaMA-2 sebelum mengimplementasikan LoRA, sehingga memberikan gambaran awal mengenai kemampuan dasar model [15]. Selanjutnya, peneliti menerapkan teknik LoRA untuk mengompresi model, mengurangi dimensi parameter, dan menguji rasio kompresi yang berbeda untuk mengamati dampaknya terhadap akurasi dan kecepatan [7]. Setelah implementasi, akurasi model dievaluasi kembali dan analisis kecepatan inferensi dilakukan untuk meningkatkan efisiensi waktu pemrosesan.

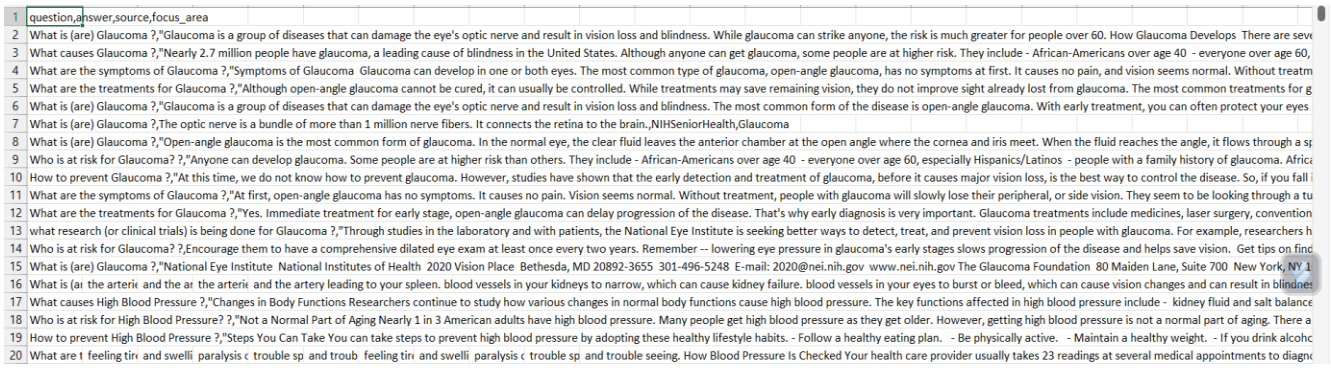
G. Integrasi *Streamlit*

Fase implementasi *Streamlit* dimulai dengan instalasi dan pengembangan skrip *Python* yang mendefinisikan antarmuka pengguna interaktif [16]. Fokus utamanya adalah menciptakan representasi yang berguna dan mudah dipahami, seperti mengintegrasikan model dan logika ilmu data. Ini membuatnya menjadi alat yang berguna untuk membagikan hasil analisis dan model secara interaktif kepada pengguna akhir [16].

III. HASIL PEMBAHASAN

A. Pengumpulan Data

Hasil dari proses *crawling data* didapatkan data sebanyak 60500 artikel dari semua artikel yang terkait dengan topik kesehatan, pendidikan, sejarah, dan teknologi dari bulan Mei 2022 sampai Januari 2024. Metode pengumpulan data penelitian ini menggunakan *web scrapping* menggunakan link postingan terkait kesehatan, pendidikan, sejarah, dan teknologi. Berikut ini merupakan hasil proses *crawling* dapat dilihat pada Gambar 4.



	question,answer,source,focus_area
2	What is (are) Glaucoma ?,"Glaucoma is a group of diseases that can damage the eye's optic nerve and result in vision loss and blindness. While glaucoma can strike anyone, the risk is much greater for people over 60. How Glaucoma Develops There are seve
3	What causes Glaucoma ?,"Nearly 2.7 million people have glaucoma, a leading cause of blindness in the United States. Although anyone can get glaucoma, some people are at higher risk. They include - African-Americans over age 40 - everyone over age 60,
4	What are the symptoms of Glaucoma ?,"Symptoms of Glaucoma Glaucoma can develop in one or both eyes. The most common type of glaucoma, open-angle glaucoma, has no symptoms at first. It causes no pain, and vision seems normal. Without treatm
5	What are the treatments for Glaucoma ?,"Although open-angle glaucoma cannot be cured, it can usually be controlled. While treatments may save remaining vision, they do not improve sight already lost from glaucoma. The most common treatments for g
6	What is (are) Glaucoma ?,"Glaucoma is a group of diseases that can damage the eye's optic nerve and result in vision loss and blindness. The most common form of the disease is open-angle glaucoma. With early treatment, you can often protect your eyes
7	What is (are) Glaucoma ?,The optic nerve is a bundle of more than 1 million nerve fibers. It connects the retina to the brain.,NIHSeniorHealth,Glaucoma
8	What is (are) Glaucoma ?,"Open-angle glaucoma is the most common form of glaucoma. In the normal eye, the clear fluid leaves the anterior chamber at the open angle where the cornea and iris meet. When the fluid reaches the angle, it flows through a sp
9	Who is at risk for Glaucoma? ?,"Anyone can develop glaucoma. Some people are at higher risk than others. They include - African-Americans over age 40 - everyone over age 60, especially Hispanics/Latinos - people with a family history of glaucoma. Africa
10	How to prevent Glaucoma ?,"At this time, we do not know how to prevent glaucoma. However, studies have shown that the early detection and treatment of glaucoma, before it causes major vision loss, is the best way to control the disease. So, if you fall i
11	What are the symptoms of Glaucoma ?,"At first, open-angle glaucoma has no symptoms. It causes no pain. Vision seems normal. Without treatment, people with glaucoma will slowly lose their peripheral, or side vision. They seem to be looking through a tu
12	What are the treatments for Glaucoma ?,"Yes. Immediate treatment for early stage, open-angle glaucoma can delay progression of the disease. That's why early diagnosis is very important. Glaucoma treatments include medicines, laser surgery, convention
13	what research (or clinical trials) is being done for Glaucoma ?,"Through studies in the laboratory and with patients, the National Eye Institute is seeking better ways to detect, treat, and prevent vision loss in people with glaucoma. For example, researchers h
14	Who is at risk for Glaucoma? ?,Encourage them to have a comprehensive dilated eye exam at least once every two years. Remember -- lowering eye pressure in glaucoma's early stages slows progression of the disease and helps save vision. Get tips on find
15	What is (are) Glaucoma ?,"National Eye Institute National Institutes of Health 2020 Vision Place Bethesda, MD 20892-3655 301-496-5248 E-mail: 2020@nei.nih.gov www.nei.nih.gov The Glaucoma Foundation 80 Maiden Lane, Suite 700 New York, NY 1
16	What is (ar the arteri and the ar the arteri and the artery leading to your spleen. blood vessels in your kidneys to narrow, which can cause kidney failure. blood vessels in your eyes to burst or bleed, which can cause vision changes and can result in blindnes
17	What causes High Blood Pressure ?,"Changes in Body Functions Researchers continue to study how various changes in normal body functions cause high blood pressure. The key functions affected in high blood pressure include - kidney fluid and salt balance
18	Who is at risk for High Blood Pressure? ?,"Not a Normal Part of Aging Nearly 1 in 3 American adults have high blood pressure. Many people get high blood pressure as they get older. However, getting high blood pressure is not a normal part of aging. There a
19	How to prevent High Blood Pressure ?,"Steps You Can Take You can take steps to prevent high blood pressure by adopting these healthy lifestyle habits. - Follow a healthy eating plan. - Be physically active. - Maintain a healthy weight. - If you drink alcohc
20	What are t feeling tir and swelli paralysis c trouble sp and troub feeling tir and swelli paralysis c trouble sp and trouble seeing. How Blood Pressure Is Checked Your health care provider usually takes 23 readings at several medical appointments to diagnc

Gambar. 4. Hasil Pengumpulan Data

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

B. *Preprocessing*

Preprocessing data dilakukan dengan menggunakan beberapa teknik, seperti *text cleaning, padding, trunncation, dan embedding.* Pemrosesan awal data tetap menjadi langkah penting dalam persiapan data untuk model pembelajaran mesin, seperti dalam konteks penerapan adaptasi peringkat rendah (LoRA) untuk efisiensi model bahasa skala besar (LLM). Langkah pra-pemrosesan ini selanjutnya bertujuan untuk mengoptimalkan kualitas dan kelengkapan data sebelum diolah oleh model. Dalam aplikasi LoRA untuk LLM, beberapa teknik pra-pemrosesan yang terlibat mencakup sanitasi teks untuk menghilangkan karakter dan gangguan asing dari teks. Data teks kemudian di isi dan dipangkas untuk mengimbangi panjangnya guna memastikan representasi data yang konsisten sebelum proses penyesuaian model dengan LoRA. Penerapan teknik *embedding* merupakan bagian penting dari preprocessing konteks LLM, dimana teks diubah menjadi representasi vektor numerik. Hal ini memungkinkan model untuk memahami dan memproses makna konten secara efektif bahkan setelah proses adaptasi menggunakan teknologi LoRA. Oleh karena itu, pemrosesan awal data dalam implementasi LoRA untuk meningkatkan efisiensi LLM memainkan peran penting dalam memastikan bahwa masukan yang diberikan ke model bersih, konsisten, dan dapat diproses secara efisien sebagai respons terhadap perubahan. Berikut ini merupakan tahapan yang dilakukan selama proses *preprocessing*.

1) *Text Cleaning*

Text Cleaning merupakan proses untuk menghapus karakter yang dinilai tidak penting, mencakup tanda baca, angka, *url*, *emoticon*, dan *hashtag* [17]. Hasil tahap *text cleaning* dapat dilihat pada Tabel I.

TABEL I
Text Cleaning

Artikel (sebelum	Text Cleaning
What's Glaucoma? Glaucoma is a group of eye conditions that can damage the optic nerve which carries visual information from the eye to the brain It is often associated with increased pressure inside the eye, which can lead to vision loss and even Blindness if left Untreated.http//www.deschope.com	*What is Glaucoma?* *Glaucoma is a group of eye conditions that can damage the optic nerve, which carries visual information from the eye to the brain. It is often associated with increased pressure inside the eye, which can lead to vision loss and even blindness if left untreated.*
What is Osteoarthritis *Affects Many Older People Osteoarthritis is the most common form of arthritis among older people, and it is one of the most frequent causes of physical disability among older adults. The disease affects both men and women. Before age 45, osteoarthritis is more common in men than in women. After age 45, osteoarthritis is more common in women. It is estimated that 33.6% (12.4 million) of individuals age 65 and older are affected by the disease.Source:www.healthscience.com/blog*	*What is Osteoarthritis?* *Affects Many Older People Osteoarthritis is the most common form of arthritis among older people, and it is one of the most frequent causes of physical disability among older adults. The disease affects both men and women. Before age 45, osteoarthritis is more common in men than in women. After age 45, osteoarthritis is more common in women. It is estimated that 33.6% (12.4 million) of individuals age 65 and older are affected by the disease.*

2) *Padding*

Padding merupakan Padding memainkan peran penting dalam fase prapemrosesan model. Fungsi utamanya adalah membuat semua teks seragam panjangnya [18]. Peneliti dapat memastikan bahwa setiap sampel data memiliki panjang yang sama dengan menambahkan token khusus dan menyematkan teks pendek. Hal ini penting untuk mengaktifkan pemrosesan data secara *batch*, yang merupakan metode umum saat melatih model berbasis transformator. Selain itu, padding juga dapat membantu mengoptimalkan pemanfaatan sumber daya komputasi, karena memungkinkan penggunaan memori GPU atau TPU secara efisien. Memastikan panjang yang seragam memungkinkan seluruh batch diproses secara paralel, sehingga meningkatkan efisiensi pemrosesan. Oleh karena itu, *padding* memainkan peran penting dalam menyiapkan data untuk melatih model bahasa besar dan memastikan konsistensi, efisiensi, dan integritas data yang diperlukan untuk hasil yang optimal [18].

TABEL II
PADDING

Text Cleaning	Padding
What is Glaucoma? *Glaucoma is a group of eye conditions that can damage the optic nerve, which carries visual information from the eye to the brain. It is often associated with increased pressure inside the eye, which can lead to vision loss and even blindness*	[PAD] [PAD]

if left untreated.	[PAD] [PAD]
What is Osteoarthritis? *Affects Many Older People Osteoarthritis is the most common form of arthritis among older people, and it is one of the most frequent causes of physical disability among older adults. The disease affects both men and women. Before age 45, osteoarthritis is more common in men than in women. After age 45, osteoarthritis is more common in women. It is estimated that 33.6% (12.4 million) of individuals age 65 and older are affected by the disease.*	[PAD] [PAD]

3) *Truncation*

Truncation ialah teknik yang digunakan dalam bidang pemrosesan bahasa alami (NLP) untuk melakukan praproses teks. Tujuan utama pemotongan adalah untuk memotong teks menjadi beberapa bagian yang lebih pendek [19]. Hal ini biasanya disebabkan oleh keterbatasan panjang input yang dapat ditangani oleh model NLP tertentu. Dengan memotong teks, Peneliti dapat memastikan bahwa teks tersebut sesuai dengan batas panjang yang dapat ditangani model [19]. Selain itu, pemotongan juga meningkatkan efisiensi komputasi karena hanya sebagian teks asli yang diproses. Hal ini sangat berguna dalam situasi dimana sumber daya komputasi terbatas [19]. Proses ini dapat dilihat pada Tabel III.

TABEL III
TRUNCATION

Padding	Truncation
Glaucoma, a type of eye disease, is a serious threat to vision as it damages the optic nerve that transmits visual signals from the eye to the brain. The condition is often accompanied by increased intraocular pressure. This is a condition where the pressure inside the eye increases beyond normal levels. If not controlled, this increased pressure can lead to visual impairment and even complete blindness.	*Glaucoma is a group of eye conditions that can damage the optic nerve, which carries visual information from the eye to the brain. It is often associated with increased pressure inside the eye, which can lead to vision loss and even blindness if left untreated.*
Osteoarthritis, which is common in the elderly, is a major form of arthritis and has a significant impact on the elderly, often resulting in disability in this population.This degenerative joint disease has no gender and affects both men and women. Surprisingly, before the age of 45, men are more likely to suffer from osteoarthritis than women. However, from that age limit, the trend reverses and women are more susceptible to the disease. Surprisingly, statistics show that 33.6% of people over the age of 65 (equivalent to 12.4 million people) suffer from the effects of osteoarthritis, thus increasing the prevalence of osteoarthritis and its serious impact on the elderly.	*Affects Many Older People Osteoarthritis is the most common form of arthritis among older people, and it is one of the most frequent causes of physical disability among older adults. The disease affects both men and women. Before age 45, osteoarthritis is more common in men than in women. After age 45, osteoarthritis is more common in women. It is estimated that 33.6% (12.4 million) of individuals age 65 and older are affected by the disease.*

4) *Embedding*

Embedding adalah representasi vektor dari data yang digunakan dalam berbagai pemrosesan bahasa. Fungsinya untuk mengubah data teks menjadi format numerik yang dapat dipahami oleh model pembelajaran mesin. *Embedding* biasanya digunakan untuk mewakili kata-kata dalam teks. Setiap kata direpresentasikan sebagai vektor numerik dengan dimensi tertentu dalam ruang vektor [20]. Dengan menggunakan teknik *Word2Vec* guna menghasilkan reprentasi numerik dari kata-kata ke dalam bentuk vektor yang dapat dimengerti

oleh komputer [20]. Representasi ini memungkinkan model untuk memahami dan memproses hubungan antar kata dalam teks, seperti persamaan dan perbedaan makna antar kata. Proses normalisasi dapat dilihat Tabel IV.

TABEL IV
EMBEDDING

Kata	Embedding
Glaucoma	[0.5, -0.3, 0.5, 0,7, 05]
Is	[0.1, 0.4, -0.2, 06]
Conditions	[0.2, 0.3, -0.4, 0.7, 0.8]
Damage	[0.4, 0.2, -0.2]
The	[0.3, -0.4, 0.3]
Optic	[0.1, 0.3, -0.5]
Nerve	[-0.3, 0.1, 0.4]
From	[-0.1, 0.5, -0.2]
Osteoarthritis	[0.4, -0.2, 0.4]
Most	[0.2, -0.1, 0.3]
Common	[0.3, -.0.2, 0.4]
People	[0.1, -0.1, 0.2]
Older	[-0.2, 0.3, 0.4]
Causes	[0.4, -0.1, 0.5]
Physical	[0.2, -0.2, 0.1]
Disability	[0.5, -0.3, 0.4]
Before	[0.3, 0.1, -0.3]
Age	[0.4, -0.2, 0.4]
Than	[0.3, 0.1, 0.2]

C. Arsitektur *Transformer*

Penerapan arsitektur Transformer bersamaan dengan penerapan adaptasi tingkat rendah (LoRA) untuk meningkatkan efisiensi model bahasa skala besar (LLM) memiliki banyak implikasi penting. Dengan menerapkan LoRA, peneliti dapat mengurangi dimensi parameter model secara signifikan, menciptakan model yang lebih ringan dan efisien tanpa mengurangi kemampuan pemahaman bahasa [10]. Pengurangan dimensi parameter ini juga berkontribusi pada peningkatan kecepatan inferensi, sehingga model dapat memberikan jawaban lebih cepat, yang sangat relevan untuk situasi yang memakan waktu [15]. Selain itu, model yang lebih efisien dapat digunakan untuk mengoptimalkan penggunaan sumber daya komputasi, memberikan manfaat tambahan dalam lingkungan dengan sumber daya terbatas. Model yang diadaptasi LoRA juga memiliki toleransi yang lebih baik terhadap fluktuasi data dan dapat menyeimbangkan akurasi dan efisiensi pemrosesan.

D. Klasifikasi Model LlaMA 2-7B

Dalam perbandingan dengan penelitian sebelumnya, penerapan LoRA pada model LlaMA 2-7B menggunakan metode yang diterapkan untuk meningkatkan efisiensi dan kecepatan infrensi pada model bahasa besar seperti penggunaan adapters dan pruning. Adapters merupakan penambahan modul kecil ke dalam modul dasar untuk memodifikasi sebagian kecil parameter agar lebih cepat dan efisien. *Pruning* melibatkan penghapusan parameter yang tidak signifikan dari model [21]. Melihat kekurangan pada metode adapters dan prunning, peneliti memilih menggunakan penerapan LoRA memiliki keuntungan dalam meningkatkan efisiensi tanpa mengorbankan akurasi. Hasil dari pelatihan model LLaMA2 merupakan momen penting dalam pengembangannya. Dalam rangkaian 200 langkah, nilai fungsi kerugian pada langkah terakhir dapat stabil pada 1,79 yang menunjukkan efektivitas proses pelatihan. Total waktu pelatihan adalah 2467,72 detik, dan kecepatan sekitar 5,19 sampel per detik memungkinkan model memproses data secara efisien. Pelatihan mencapai sekitar 78% dari *target epoch*, namun evaluasi lebih lanjut diperlukan untuk memahami konvergensi dan keakuratan model pada *dataset*. Kompleksitas model ini tercermin dari total *FLOP* yang berjumlah sekitar 4,78 triliun, yang menunjukkan besarnya jumlah operasi yang diperlukan. Model LLaMA2 menunjukkan peningkatan yang signifikan dalam efisiensi klasifikasi teks generatif setelah menerapkan adaptasi peringkat rendah (LoRA),

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

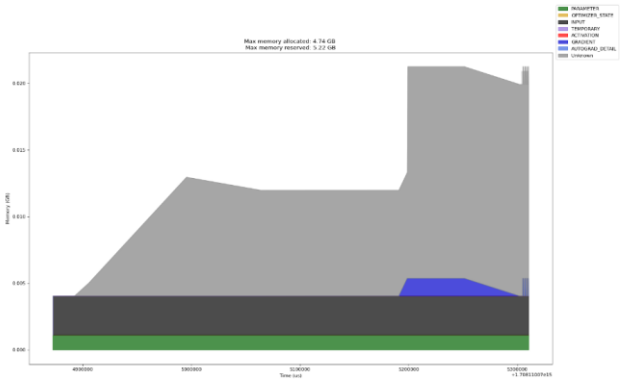
terutama saat menjawab pertanyaan kesehatan [14]. Model ini berfokus pada pertanyaan kesehatan dan dilatih untuk menghasilkan teks respons yang dikategorikan berdasarkan kategori topik kesehatan yang relevan [22]. Model ini dapat mencapai tingkat akurasi yang serupa dengan model yang lebih besar dengan menggunakan sumber daya yang lebih sedikit. Potensi model sebagai alat klasifikasi teks generatif yang efektif untuk masalah kesehatan ditunjukkan dengan hasil yang mengesankan setelah penyesuaian dan pelatihan. Hasil klasifikasi LlaMA 2-7B dapat dilihat pada Tabel V.

TABEL V
KLASIFIKASI MODEL

Pertanyaan	Aktual	Prediksi
What is Glaucoma?	*Glaucoma is a group of eye conditions that can damage the optic nerve. It is often associated with increased pressure inside the eye, which can lead to vision loss and even blindness if left untreated.*	*Glaucoma is a group of eye conditions that can damage the optic nerve. It is associated with increased pressure inside the eye and can lead to vision loss and blindness if untreated.*
What is Osteoathritis?	*Osteoarthritis is the most common form of arthritis among older people. It causes physical disability and affects both men and women. Before age 45, it is more common in men, but after age 45, it is more common in women. Around 33.6% (12.4 million) of individuals aged 65 and older are affected by this disease.*	*Osteoarthritis is the most common form of arthritis among older people. It causes physical disability and affects both men and women. Before age 45, it is more common in men, but after age 45, it is more common in women. Around 33.6% (12.4 million) of individuals aged 65 and older are affected by this disease.*

E. Implementasi LoRA

Penerapan adaptasi peringkat rendah (LoRA) pada model LlaMA 2-7B terbukti memberikan dampak positif terhadap efisiensi dan kecepatan inferensi. LoRA, yang merupakan singkatan *Low Rank Adaptation*, bertujuan mengurangi dimensi parameter model dasar dan skala besar akan meningkatkan efisiensi dengan rasio kompresi 2x, 4x, dan 8x, sehingga memungkinkan fleksibilitas dalam menyesuaikan keseimbangan antara akurasi dan kecepatan. Prinsip dibalik LoRA melibatkan matriks dekomposisi peringkat rendah ke setiap lapisan dalam arsitektur transfromer, yang memungkinkan model untuk mengangkap parameter yang paling berpengaruh. Penelitian sebelumnya, LoRA telah terbukti meningkatkan efisiensi dan kecepatan infrensi pada berbagai tugas pemrosesan bahasa. Dibandingkan metode sebelumnya, tanpa menerapkan LoRA beban memori dan GPU yang digunakan sangat tinggi dan membutuhkan biaya yang mahal [23]. Dengan menerapkan LoRA pada model LlaMA2-7B yang menggabungkan quantization dan penggunaan GPU virtual dengan spesifikasi GPU V100 dan P100 yang dapat berjalan hingga 146% dan 63% lebih cepat sehingga menghasilkan model adaptif yang cocok untuk tugas dengan respon yang cepat dan hemat biaya [24]. Peningkatan kecepatan inferensi yang signifikan, meskipun kehilangan akurasinya minimal, merupakan bukti bahwa implementasi LoRA berhasil mencapai efisiensi tanpa mengorbankan kekuatan klasifikasi. Model LoRA-LlaMA 2-7B berjalan 2x, 4x, dan 8x lebih cepat dibandingkan model dasar, memberikan solusi efisien terkait respon cepat dalam lingkungan aplikasi yang memerlukan kinerja real-time. Oleh karena itu, penerapan LoRA ke dalam model LlaMA 2-7B memberikan solusi efektif untuk meningkatkan efisiensi model tanpa mempengaruhi kemampuan klasifikasi secara signifikan. Dalam situasi di mana kecepatan inferensi merupakan prioritas, model adaptif LoRA dapat memberikan solusi optimal untuk menangani tugas respons cepat di berbagai aplikasi pemrosesan bahasa. Dibawah ini merupakan hasil implementasi LoRA dapat dilihat pada Gambar 5.



Gambar. 5. Visualisasi Pemakaian GPU

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 1881-1890

Dari visualisasi tersebut, terlihat bahwa memori GPU secara signifikan berkurang ketika model LlaMA2 dilatih dengan menerapkan LoRA dibandingkan dengan pelatihan tanpa LoRA. Hal ini menandakan bahwa LoRA dapat meningkatkan efisiensi LLM dengan mengurangi penggunaan sumber daya komputasi.

F. Evaluasi Model

Hasil evaluasi model LlaMA 2-7B dan penerapan *low-rank adaptation* (LoRA) pada 200 langkah pelatihan menunjukkan keberhasilan konvergensi dengan nilai *loss function* stabil sebesar 1,79. Meskipun bukti positif telah diberikan untuk efektifitas pelatihan, evaluasi lebih lanjut terhadap metrik validasi seperti presisi dan perolehan kembali diperlukan untuk memahami performa model secara keseluruhan. Waktu pelatihan yang singkat sekitar 2467,72 detik dan kecepatan pemrosesan data yang sangat baik sekitar 5187 sampel per detik mencerminkan efisiensi pelatihan, didukung oleh total sekitar 4,78 triliun *FLOP*, dan efisiensi operasional model menunjukkan kompleksitas. Membandingkan penggunaan memori GPU model LlaMA2 dengan LoRA, peneliti menemukan bahwa implementasi LoRA model LlaMA2 dengan 45,6 miliar parameter pelatihan hanya memerlukan memori GPU sebesar 4,74 GB. Kesimpulannya, penerapan LoRA di LlaMA2 secara signifikan menyederhanakan penggunaan memori GPU tanpa mengurangi performa pelatihan, sehingga menunjukkan potensi model untuk digunakan secara lebih efektif di berbagai lingkungan. Berikut ini adalah hasil evaluasi model yang dapat dilihat pada Tabel V.

TABEL V
EVALUASI MODEL

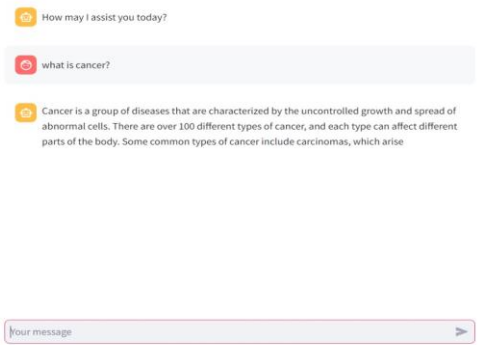
Model	Parameter Pelatihan	Memori GPU
LlaMA2 *With* LoRA	4.56080	4.74

Pada tabel IV menunjukkan perbandingan penggunaan memori GPU antara dua model yaitu LlaMA2 dengan LlaMA2 selama proses pelatihan. Model LlaMA2 dengan menerapkan LoRA memiliki sekitar 45.608 juta parameter pelatihan dan hanya memerlukan memori GPU sekitar 4.74 GB. Disisi lain, model LlaMA 2-7B tanpa menggunakan LoRA mencapai 38.9004 ratus juta parameter pelatihan dan memerlukan alokasi memori GPU yang lebih banyak sekitar 23,1 GB.

G. Integrasi *Streamlit*

Pada tahap ini, implementasi *streamlit* model LlaMA 2-7B, yang diimplementasikan *menggunakan Low-Rank Adaptation* (LoRA), memberikan solusi antarmuka pengguna yang sederhana dan efisien. *Streamlit* memungkinkan pengguna berinteraksi dengan model tanpa memerlukan pengetahuan teknis yang mendalam. Antarmuka yang bersih dan intuitif memungkinkan pengguna memasukkan pertanyaan dan teks dengan cepat serta mendapatkan hasil instan. Interaktivitas *streamlit* memungkinkan pengguna menjelajahi fitur LlaMA Model 2 secara dinamis. Integrasi model *streamlit* dan LlaMA 2-7B yang menggunakan teknologi LoRA merupakan aspek penting dalam implementasi ini. Dengan memanfaatkan model terkompresi, antarmuka memungkinkan respons inferensi yang lebih cepat dan meningkatkan pengalaman pengguna saat menggunakan model.

Selain itu, Pengguna dapat dengan mudah menyesuaikan antarmuka sesuai dengan kebutuhan proyek dan mengelola serta mengembangkan fungsionalitas tambahan secara efisien. Oleh karena itu, implementasi model LlaMA 2-7B menggunakan LoRA dengan *streamlit* tidak hanya meningkatkan aksesibilitas model oleh pengguna non-teknis, namun juga menyediakan antarmuka yang responsif dan mudah beradaptasi dengan kebutuhan proyek. Hasil dari proses implementasi *streamlit* dapat dilihat pada Gambar 6.



Gambar. 6. Visualisasi Implementasi *Streamlit*

IV. KESIMPULAN

Berdasarkan hasil yang telah didapat, penerapan adaptasi tingkat rendah (LoRA) pada model LlaMA 2-7B menunjukkan keberhasilan besar dalam meningkatkan efisiensi model bahasa skala besar (LLM). Model LoRA-LlaMA 2-7B mencapai kompresi parameter yang signifikan, menciptakan model yang lebih efisien secara komputasi. Model LoRA-LlaMA 2-7B berhasil mencapai kompresi parameter yang signifikan, menghasilkan model yang lebih efisien secara komputasi dengan hanya memerlukan sekitar 4.74 GB memori GPU. Untuk lebih meningkatkan implementasi lebih lanjut, peneliti menyarankan untuk pengujian mendalam terhadap dampak rasio kompresi pada performa model, serta menambahkan data yang lebih kompleks.

DAFTAR PUSTAKA

[1] F. Petroni *et al.*, "Language models as knowledge bases?," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 2463–2473, 2019, doi: 10.18653/v1/d19-1250.
[2] D. Chenxi, "How to Build an AI Tutor that Can Adapt to Any Course and Provide Accurate Answers Using Large Language Model and Retrieval-Augmented Generation," pp. 1–9.
[3] H. Azzuni, S. Jamal, and A. Elsaddik, "uTalk: Bridging the Gap Between Humans and AI," no. 1, pp. 12–15, 2023, [Online]. Available: http://arxiv.org/abs/2310.02739
[4] E. Hu *et al.*, "Lora: Low-Rank Adaptation of Large Language Models," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, pp. 1–26, 2022.
[5] S. Sun, D. Gupta, and M. Iyyer, "Exploring the impact of low-rank adaptation on the performance , efficiency , and regularization of RLHF," pp. 1–14, 2023.
[6] T. For, "L O RA-FA : M EMORY - EFFICIENT L OW - RANK A DAPTA -," vol. 1, pp. 1–15, 2023.
[7] Y. Li *et al.*, "LoftQ : LoRA-Fine-Tuning-Aware Quantization for Large," vol. 2023, 2023.
[8] J. Armengol-Estapé, "A pipeline for large raw text preprocessing and model training of language models at scale," 2021.
[9] J. Vig, "Analyzing the Structure of Attention in a Transformer Language Model," 2019.
[10] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, "Larger-Scale Transformers for Multilingual Masked Language Modeling," *RepL4NLP 2021 - 6th Work. Represent. Learn. NLP, Proc. Work.*, pp. 29–33, 2021, doi: 10.18653/v1/2021.repl4nlp-1.4.
[11] P. Gao *et al.*, "LLaMA-Adapter V2 : Parameter-Efficient Visual Instruction Model".
[12] R. Zhang *et al.*, "LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention," pp. 1–22, 2023, [Online]. Available: http://arxiv.org/abs/2303.16199
[13] Y. Bai *et al.*, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," 2022, [Online]. Available: http://arxiv.org/abs/2204.05862
[14] S. Lermen, C. Rogers-smith, and J. Ladish, "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," 2022.
[15] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing," pp. 1–42, 2021, [Online]. Available: http://arxiv.org/abs/2108.05542
[16] J. M. Nápoles-Duarte, A. Biswas, M. I. Parker, J. P. Palomares-Baez, M. A. Chávez-Rojo, and L. M. Rodríguez-Valdez, "Stmol: A component for building interactive molecular visualizations within streamlit web-applications," *Front. Mol. Biosci.*, vol. 9, no. September, pp. 1–10, 2022, doi: 10.3389/fmolb.2022.990846.
[17] M. Saravanan, P. C. R. Raj, and S. Raman, "Summarization and categorization of text data in high-level data cleaning for information retrieval," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 461–474, 2003, doi: 10.1080/713827177.
[18] A. Nagarajan and A. Raghunathan, "TokenDrop + BucketSampler: Towards Efficient Padding-free Fine-tuning of Language Models," *Find. Assoc. Comput. Linguist. EMNLP 2023*, pp. 11682–11695, 2023, doi: 10.18653/v1/2023.findings-emnlp.782.
[19] J. Hewitt, C. D. Manning, and P. Liang, "Truncation Sampling as Language Model Desmoothing," *Find. Assoc. Comput. Linguist. EMNLP 2022*, no. 1, pp. 3414–3427, 2022, doi: 10.18653/v1/2022.findings-emnlp.249.
[20] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," no. 2, pp. 1–5, 2014, [Online]. Available: http://arxiv.org/abs/1402.3722
[21] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A Simple and Effective Pruning Approach for Large Language Models," pp. 1–22, 2023, [Online]. Available: http://arxiv.org/abs/2306.11695
[22] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023, [Online]. Available: http://arxiv.org/abs/2307.09288
[23] T. Dettmers, "QL O RA : Efficient Finetuning of Quantized LLMs," no. NeurIPS, 2023.
[24] S. Li *et al.*, "CaraServe: CPU-Assisted and Rank-Aware LoRA Serving for Generative LLM Inference," 2024, [Online]. Available: http://arxiv.org/abs/2401.11240