COMPRESSION CHARACTERISTIC OF LLM USING PRUNING METHODS ON EDGE DEVICE

SKRIPSI

Submitted to fulfill one of the requirements for obtaining a Bachelor's degree
Informatics Study Program

prepared by
SULTAN GEMILANG KEMADI
21.11.4303

To

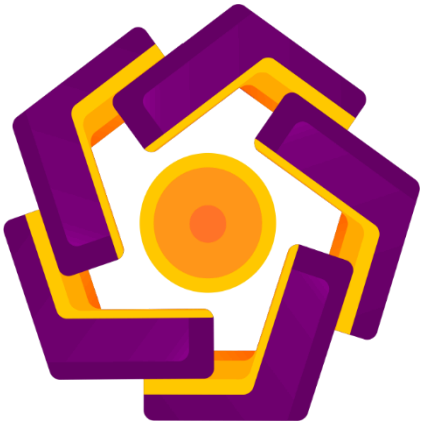
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2025

COMPRESSION CHARACTERISTIC OF LLM USING PRUNING METHODS ON EDGE DEVICE

SKRIPSI

Submitted to fulfill one of the requirements for obtaining a Bachelor's degree

Informatics Study Program



prepared by

SULTAN GEMILANG KEMADI

21.11.4303

To

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

SKRIPSI

COMPRESSION CHARACTERISTIC OF LLM USING PRUNING METHODS ON EDGE DEVICE

yang disusun dan diajukan oleh

Sultan Gemilang Kemadi

21.11.4303

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 28 Juli 2025

Dosen Pembimbing,

Arief Setyanto, S.Si., MT., Ph.D
NIK. 190302036

HALAMAN PENGESAHAN

SKRIPSI

COMPRESSION CHARACTERISTIC OF LLM USING PRUNING METHODS ON EDGE DEVICE

yang disusun dan diajukan oleh

Sultan Gemilang Kemadi

21.11.4303

Telah dipertahankan di depan Dewan Penguji
pada tanggal 28 Juli 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Theopilus Bayu Sasongko, S.Kom., M.Eng. **NIK. 190302375**	
Arifiyanto Hadinegoro, S.Kom., M.T. **NIK. 190302289**	
Arief Setyanto, S.Si., M.T., Ph.D. **NIK. 190302036**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 28 Juli 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Sultan Gemilang Kemadi
NIM : 21.11.4303

Menyatakan bahwa Skripsi dengan judul berikut:

COMPRESSION CHARACTERISTIC OF LLM USING PRUNING METHODS ON EDGE DEVICE

Dosen Pembimbing : Arief Setyanto, S.Si., MT., Ph.D

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 28 Juli 2025

Yang Menyatakan,

Sultan Gemilang K

MOTTO

"Youth is not a time of life; it is a state of mind; . . . a vigor of the emotions; it is the freshness of the deep springs of life . . . . Nobody grows old merely by a number of years. We grow old by deserting our ideals."

-Samuel Ullman, "Youth", 1918

HALAMAN PERSEMBAHAN

Dengan puji syukur atas kesempatan yang diberikan Allah SWT, saya telah menyelesaikan skripsi yang telah dikerjakan selama jangka waktu 3 bulan dengan data seadanya. Dengan ini, saya persembahkan skripsi ini kepada:

1. Kedua orangtua tercinta saya, yang telah terus menyemangati, mengingat, dan juga mendukung skripsi saya walaupun berjarak 3.417 KM ke arah timur laut.
2. Kakak saya, Azzallie Sedayu Putri Kemadi yang telah membantu menyemangati dan mendanai beberapa proyek yang saya lakukan.
3. Bapak Arief Setyanto, S.Si., MT., Ph.D, selaku dosen pembimbingan skripsi dan Garuda Ace yang selalu memberikan ide dan solusi di berbagai permasalahan penelitian.
4. Dosen wali ibu Nuraini, M.Kom. yang telah memandu saya selama 4 tahun terakhir di Univeristas AMIKOM Yogyakarta.
5. Teman sekelas saya Ahmad Hawari, Putu Prima Wahyutama Sutaya, Widi Dwiky Alhamdi, Dhuta Azikirra Subroto, Iqbal Risqi Saputra, Wildansyah Maparoja, Pramudya Anggara Zitha, dan semua teman kelas 21-IF-07 yang telah mengajak bermain dan menemani selama 4 tahun terakhir.
6. Anggota Telad-an HUT, Hafizh Favian Setiadhy Pratama, Faris Firman Saputra, Wahyu Bimo Arianto, dan anggota lainnya yang telah menemani saya pada pembicaraan tengah malam yang menarik dan menyenangkan.
7. Anggota Lycoreco, Daffa Aunillah, Muhammad Khadafie Satya Sudarto, Muhammad Aden Misbah, Celvin Tetiandra Suroso, dan anggota lainnya yang telah mengisi waktu luang menjadi sesuatu yang produktif dan informatif selama beberapa tahun ini.
8. Policepicadilly (ポリスピカデリー), ASU (明透), Kenshi Yonezu, PinocchioP, DECO*27, Nyarons, DAZBEE, Koresawa (コレサワ), artist-artist Jpop lainya yang telah menemani malam lembur skripsi.
9. Semua orang yang tidak bisa saya sebutkan pada halaman persembahan skripsi ini.

FOREWORD

With gratitude and praise to Allah SWT for the opportunity granted, and for the blessings and grace bestowed, the author has been able to complete this thesis entitled “Compression Characteristic of LLM using Pruning Methods on Edge Device.”

This thesis discusses the characteristics of latency, performance, and power in LLM pruning implemented on edge devices. The study is based on the growing complexity of LLMs and the increasing demand for edge devices capable of running AI.

The completion of this thesis would not have been possible without the help, support, and input from various parties. Therefore, the researcher would like to express the deepest gratitude to:

1. Both of my parents, Mr. Alif Kemadi and Mrs. Nurhayati Afiyah, thank you for all of the love and support, prayer, and motivation. I hope I can make you proud one day. Thank you.
2. My older sister, whom I adore, thank you for the support of all the encouragement, and the mischief that happened between us. I hope we can finally conquer the world and become successful people in the future together.
3. Mr. Arief Setyanto, S.Si., MT., Ph.D, as my thesis advisor who guides me from the Garuda Ace program and finally my thesis. Thank you for guiding me patiently along the course of the Garuda Ace program and my thesis assignment.
4. Mrs. Nuraini, M.Kom., as my academic advisor, who guided me from the beginning of the semester, and this final thesis.
5. Dr. In Kee Kim, Rakandhiya D. Rachmanto, Zaki Sukma, as part of the Garuda Ace program, Thank You for giving me the opportunity to work with you. Without it, this research would not have existed.

6. All professors, lecturers, teachers, and staff of the University of AMIKOM Yogyakarta who always help, guide, and provide all the necessities that I need to complete my study
7. My classmates from 21-IF-07, Ahmad Hawari, Putu Prima Wahyutama Sutaya, Widi Dwiky Alhamdi, Dhuta Azikirra Subroto, Iqbal Risqi Saputra, Wildansyah Maparoja, Pramudya Anggara Zitha, and all those I didn't mention, who make my university days much more enjoyable.
8. Telad-an HUT group members, Hafizh Favian Setiadhy Pratama, Faris Firman Saputra, Wahyu Bimo Arianto, and other members who accompany me with stupid and funny talks at the very late talks session.
9. Lycoreco family members, Daffa Aunillah, Muhammad Khadafie Satya Sudarto, Muhammad Aden Misbah, Celvin Tetiandra Suroso, and other members, for the amazing trips and adventures that we went through.
10. Talented Japanese music artists such as Policepicadilly (ポリスピカデリー), ASU (明透), Kenshi Yonezu, PinocchioP, DECO*27, Nyarons, indigo la End, DAZBEE, Koresawa (コレサワ), and many other J-pop artists' songs have accompanied the late nights working on my thesis.
11. And last but not least, for those who read this paper and for those whom I cannot mention in this section.

The author is aware that this thesis is far from perfect because Allah is the only one who has possession of perfection. Because of this, I sincerely embrace all criticism and suggestions to improve this thesis. In the present or in the future, I hope that my thesis will be beneficial to those who are interested.

Last but not least, forgive all of my errors in case there are still any incorrect or insufficient words in my thesis.

Yogyakarta, 28 Juli 2025

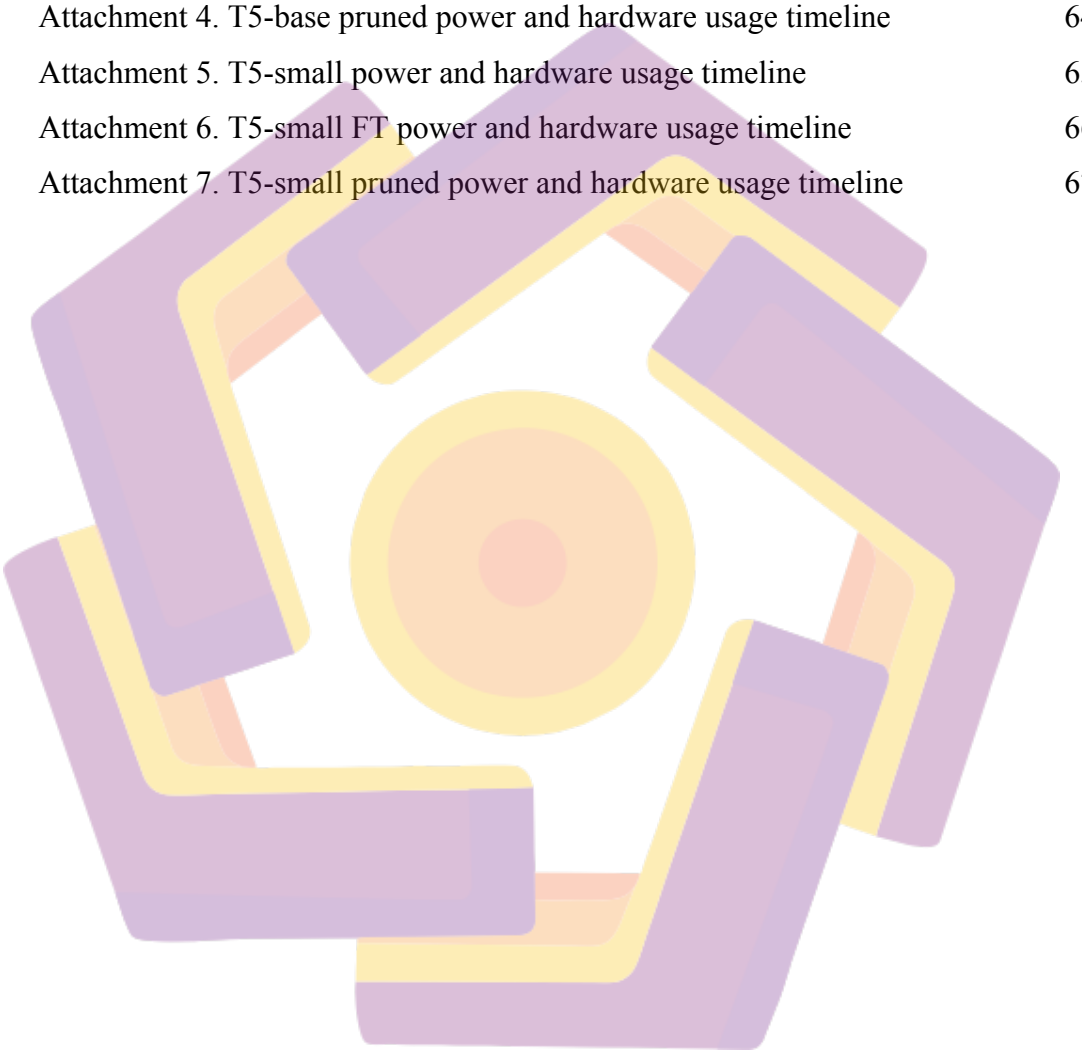
Author

TABLE OF CONTENTS

TABLE LIST

FIGURE LIST

ATTACHMENT LIST



SYMBOL AND ABBREVIATION LIST

SOTA	State-Of-The-Art
LLMs	Large Language Models
SDK	Software Development Kit
CPU	Central Processing Unit
GPU	Graphics Processing Unit
iGPU	Integrated Graphics Processing Unit
RAM	Random Access Memory
iRAM	Integrated Random Access Memory
TTFT	Time To First Token
TGT	Total Generated Time
ITL	Inter-Token Latency
TPS	Token Per Second
NLP	Natural Language Processing
MMLU	Massive Multitask Language Understanding
MQA	Multi-Query Attention
MLA	Multi-head Latent Attention
NSA	Native Sparse Attention
MoE	Mixture of Experts
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Ranking Adaptation
T5	Text-To-Text Transfer Transformer
HF	HunggingFace
PC	Personal Computer
-M	-Million
GPT	Generative Pre-trained Transformer
-GB	Gigabyte
FFNs	Feed-Forward Networks
kWh	KiloWatt Hour
EDA	Exploratory Data Analys

TERM LIST

Latency	Delay between action and reaction
Edge Device	Hardware used in edge computing
Token	Basic unit of text that the model processes
float16	16-bit binary floating point number
bfloat16	Variation of float16 optimized for AI
Quantization	AI compression by reducing precision
Pruning	AI compression by removing parameters
Sparsity	Ratio of zero in tensors
Tokenization	Breaking down text into smaller units called tokens
Detokenization	Converting tokens back into human-readable text
HF Evaluate	HuggingFace model evaluation library
Fine-tuning	Training model to improve performance
Repository	A place to store files, code, etc.
Tegrastats	Utility for hardware monitoring

INTISARI

Dengan perkembangan terbaru dalam Large Language Models (LLM), kita telah menyaksikan peningkatan akurasi dan performa dalam berbagai tugas. Meskipun terjadi peningkatan tersebut, masih terdapat beberapa masalah terkait sumber daya komputasi, penggunaan memori, dan konsumsi energi. Masalah-masalah ini membatasi penggunaan praktis pada perangkat dengan sumber daya terbatas, seperti edge device. Untuk mengatasi hal ini, teknik kompresi model, khususnya metode pruning, telah mendapat perhatian sebagai solusi yang efektif. Studi ini mengeksplorasi dampak dari berbagai pendekatan pruning terhadap LLM. Melalui evaluasi eksperimental, kami meneliti karakteristik efek pruning terhadap latensi inferensi, metrik performa, dan konsumsi daya. Studi kami menyoroti tiga temuan utama: (1) Meskipun pruning dapat secara signifikan mempercepat inferensi, hasil latensi menunjukkan variasi yang besar, yang dapat menyebabkan peningkatan variabilitas dan meningkatkan latensi TTFT tergantung pada input, sehingga dapat berdampak pada responsivitas waktu nyata; (2) Pruning dapat mengurangi penggunaan perangkat keras hingga 10% dan konsumsi daya hingga 14%, yang menunjukkan komputasi yang lebih efisien; dan (3) Pruning pada model yang lebih besar lebih efektif dibandingkan dengan pruning pada model yang sudah kecil. Temuan-temuan ini diharapkan dapat memberikan wawasan berharga untuk mengoptimalkan penerapan LLM di lingkungan dengan keterbatasan sumber daya serta meningkatkan metode pruning di kedepannya.

Kata kunci: LLMs, Edge Device, Kompresi, Pruning, Karakteristik.

ABSTRACT

With recent developments in Large Language Models (LLMs), we have witnessed improved accuracy and performance in various tasks. Despite these improvements, there are issues in terms of computational resources, memory usage, and energy consumption. These issues limit the practical usage in resource-limited computers such as edge devices. To solve this, model compression techniques, particularly pruning methods, have gained attention as an effective solution for these challenges. This study explores the impact of various pruning approaches on LLMs. Through experimental evaluations, we examine the characteristic effect of pruning on inference latency, performance metrics, and power consumption. Our study highlights three insights: (1) Although pruning can significantly reduce the inference speed, the latency results show a wide variability that may lead to increased variability and increase TTFT latency depending on the input, which may impact real-time responsiveness, (2) Pruning reduces the hardware usage up to 10% and power usage up to 14%, indicating a more efficient computation, and (3) Pruning larger model is more effective compared to pruning already smaller model. These insights hope to provide valuable knowledge for optimizing LLM deployment in constrained environments and improving future pruning methods.

.

***Keyword*:** LLMs, Edge Device, Compression, Pruning, Characteristic