

LEMBAR PERSEMBAHAN

Sebuah perjalanan panjang tidak diukur dari seberapa cepat kita sampai, melainkan seberapa gigih kita menolak untuk putar balik.

Puji dan syukur yang tak terhingga kepada Allah S.W.T atas segala rahmat dan kekuatan-Nya. Skripsi ini bukan sekadar syarat kelulusan, melainkan sebuah pembuktian dan persembahan tulus untuk:

1. **Bapak dan Ibu**, dua sosok yang tak pernah lelah merapalkan doa dalam diam. Terima kasih telah memaklumi setiap waktu istirahat yang tersita, memaafkan jarak yang kadang tercipta karena kesibukan, dan menjadi alasan terbesarku untuk terus melangkah maju.
2. **Keluarga dan sosok-sosok istimewa** di sekelilingku. Terima kasih untuk telinga yang selalu sedia mendengar, ruang aman untuk berbagi keluh kesah, dan semangat yang terus disuntikkan saat semuanya terasa buntu.
3. **Teruntuk diriku sendiri**. Terima kasih sudah menolak untuk menyerah. Dari kerasnya jalanan dan debu yang menemani langkah di awal masa perkuliahan, padatnya membagi raga dengan dunia kerja, hingga malam-malam tanpa tidur demi merangkai barisan kode dan merampungkan riset ini. Lelah itu akhirnya menemukan garis akhirnya.

Karya ini ada karena doa yang tak terputus dan tekad yang menolak padam.

KATA PENGANTAR

Alhamdulillah, Dengan mengucapkan puji syukur kehadiran Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan tugas ini dengan baik. Skripsi pada Program Sarjana ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul Skripsi, yang penulis ambil sebagai berikut, **“Analisis Performa Post-Training Quantization (PTQ) pada Small Language Model untuk Implementasi Edge Computing Berbasis Android”**.

Tujuan penulisan Skripsi pada Program Sarjana ini dibuat sebagai salah satu syarat kelulusan Sarjana Universitas Bina Sarana Informatika. Sebagai bahan penulisan diambil berdasarkan hasil penelitian (eksperimen), observasi dan beberapa sumber literatur yang mendukung penulisan ini. Penulis menyadari bahwa tanpa bimbingan dan dorongan dari semua pihak, maka penulisan Skripsi ini tidak akan berjalan lancar. Oleh karena itu pada kesempatan ini, izinkanlah penulis menyampaikan ucapan terima kasih kepada:

1. Rektor Universitas Bina Sarana Informatika.
2. Dekan Fakultas Teknik dan Informatika
3. Ketua Program Studi Teknologi Informasi Universitas Bina Sarana Informatika.
4. Bapak Taufik Rahman, M.Kom selaku Dosen Pembimbing Skripsi.
5. Staff / karyawan / dosen di lingkungan Universitas Bina Sarana Informatika.
6. Orang tua tercinta yang telah memberikan dukungan moral maupun spiritual.

Serta semua pihak yang terlalu banyak untuk disebut satu persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga Skripsi ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 29 Juni 2026
Penulis



Fardani Azhar

ABSTRAK

Fardani Azhar (17220245), Analisis Performa *Post-Training Quantization* (PTQ) pada *Small Language Model* untuk Implementasi *Edge Computing* Berbasis Android

Pemanfaatan *Small Language Model* (SLM) secara *on-device* di Android sering terkendala oleh kapasitas RAM 8 GB yang dibagi-pakai dengan sistem operasi. Akibatnya, pemuatan model berpresisi penuh FP16 berisiko tinggi memicu *Out of Memory* dan *Force Close*. Penelitian ini bertujuan menganalisis performa metode *Post-Training Quantization* (PTQ) berformat GGUF *k-quants* (Q3_K_M, Q4_K_M, Q5_K_M) terhadap dua arsitektur SLM, yaitu LFM 2.5 (1,2B) dan Qwen 3.5 (2B), pada perangkat Tecno Pova 5 melalui lingkungan Termux dan *llama.cpp*. Metode yang digunakan adalah eksperimen kuantitatif komparatif. Pengujian ini mengukur reduksi ukuran berkas, konsumsi *peak* RAM, kecepatan *prompt* dan *generation*, serta meninjau degradasi kognitif melalui *Perplexity* dan evaluasi akurasi pada *benchmark* MMLU, GSM8K, HumanEval, dan MT-Bench TTR. Hasil penelitian menunjukkan varian Q4_K_M sebagai *sweet spot* terbaik. Varian ini mereduksi ukuran berkas hingga 66–68% (dari 3,6 GB menjadi 1,2 GB pada Qwen dan 2,2 GB menjadi 698 MB pada LFM) serta mengoptimalkan *peak* RAM sebesar 36,9% (dari 2.303,47 MB menjadi 1.452,97 MB pada LFM) dan 31,6% (dari 3.744,25 MB menjadi 2.562,88 MB pada Qwen). Kecepatan *generation* meningkat hingga 2,65 kali lipat (dari 1,87 t/s menjadi 4,95 t/s pada Qwen dan 5,57 t/s menjadi 13,67 t/s pada LFM) dengan retensi akurasi mencapai 88,9% (skor 112/300 vs. *baseline* 126/300 pada LFM) dan 100% (skor 119/300 vs. *baseline* 119/300 pada Qwen). Sebaliknya, Q3_K_M memicu anomali *prompt speed* dan degradasi akurasi signifikan, sehingga tidak direkomendasikan untuk produksi.

Keywords: *Post-Training Quantization, Small Language Model, Edge Computing, GGUF, k-quants, Android, Helio G99, Perplexity.*

ABSTRACT

Fardani Azhar (17220245), *Performance Analysis of Post-Training Quantization (PTQ) on Small Language Models for Android-Based Edge Computing Implementation*

The on-device deployment of Small Language Models (SLMs) on Android is frequently constrained by the 8 GB RAM capacity shared with the operating system. Consequently, loading full-precision FP16 models poses a high risk of triggering Out of Memory and Force Close errors. This study aims to analyze the performance of the Post-Training Quantization (PTQ) method utilizing the GGUF k-quants format (Q3_K_M, Q4_K_M, Q5_K_M) on two SLM architectures, namely LFM 2.5 (1.2B) and Qwen 3.5 (2B), running on a Tecno Pova 5 device through the Termux environment and llama.cpp. The method employed is a comparative quantitative experiment. The testing measures file size reduction, peak RAM consumption, prompt and generation speeds, and evaluates cognitive degradation via Perplexity and accuracy benchmarks including MMLU, GSM8K, HumanEval, and MT-Bench TTR. The findings indicate that the Q4_K_M variant acts as the optimal sweet spot. It reduces file sizes by 66–68% (from 3.6 GB to 1.2 GB on Qwen and 2.2 GB to 698 MB on LFM) and improves peak RAM efficiency by 36.9% (from 2,303.47 MB to 1,452.97 MB on LFM) and 31.6% (from 3,744.25 MB to 2,562.88 MB on Qwen). Generation speed increases by up to 2.65 times (from 1.87 t/s to 4.95 t/s on Qwen and 5.57 t/s to 13.67 t/s on LFM) with accuracy retention reaching 88.9% (score 112/300 vs. baseline 126/300 on LFM) and 100% (score 119/300 vs. baseline 119/300 on Qwen). In contrast, the Q3_K_M variant causes prompt speed anomalies and significant accuracy degradation, making it unadvisable for production.

Keywords: *Post-Training Quantization, Small Language Model, Edge Computing, GGUF, k-quants, Android, Helio G99, Perplexity.*

DAFTAR ISI

LEMBAR PERNYATAAN KEASLIAN SKRIPSI	ii
LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iii
LEMBAR PERSETUJUAN DAN PENGESAHAN SKRIPSI	iv
LEMBAR PEDOMAN PENGGUNAAN HAK CIPTA.....	v
LEMBAR KONSULTASI SKRIPSI.....	vi
LEMBAR PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
ABSTRAK	ix
ABSTRACT.....	x
DAFTAR ISI	xi
DAFTAR SIMBOL.....	xiii
DAFTAR GAMBAR	xiv
DAFTAR TABEL.....	xv
DAFTAR LAMPIRAN	xvi
BAB I PENDAHULUAN.....	17
1.1. Latar Belakang Masalah.....	17
1.2. Identifikasi Permasalahan	19
1.3. Perumusan Masalah	21
1.4. Tujuan dan Manfaat	22
1.5. Metode Penelitian.....	25
1.6. Teknik Pengumpulan Data	28
1.7. Ruang Lingkup.....	31
1.8. Hipotesis.....	34
BAB II LANDASAN TEORI	38
2.1. Tinjauan Pustaka	38
2.2. Penelitian Terkait.....	44
2.3. Tinjauan Objek Penelitian.....	48
BAB III METODOLOGI PENELITIAN.....	51
3.1. Tahapan Penelitian	51

3.2.	Instrumen Penelitian.....	55
3.3.	Metode Pengumpulan Data	56
3.4.	Metode Analisis Data	58
BAB IV HASIL PENELITIAN DAN PEMBAHASAN		61
4.1.	Lingkungan Pengujian dan Skenario	61
4.2.	Hasil Uji Efisiensi Infrastruktur (Hardware).....	61
4.3.	Hasil Uji Degradasi Kognitif (Software / AI)	67
4.4.	Pembahasan Analisis Komparatif	71
BAB V PENUTUP.....		86
5.1.	Kesimpulan	86
5.2.	Saran.....	90
DAFTAR PUSTAKA.....		94
DAFTAR RIWAYAT HIDUP.....		97
SURAT PERNYATAAN KEBENARAN/KEABSAHAN DATA HASIL RISET UNTUK KARYA ILMIAH.....		98
BUKTI HASIL PENGECEKAN PLAGIARISME		99
LAMPIRAN A SKRIP OTOMATISASI PENGUJIAN (RTX 3060)		100
LAMPIRAN B SKRIP EKSEKUSI INFERENSI EDGE (Tecno Pova 5)		132
LAMPIRAN C HASIL BENCHMARK.....		136
LAMPIRAN D LANGKAH-LANGKAH PERSIAPAN LINGKUNGAN DAN KUANTISASI MODEL		140
LAMPIRAN E DAFTAR AKRONIM DAN ISTILAH TEKNIS		144
LAMPIRAN F SCREENSHOT TERMINAL		146

DAFTAR SIMBOL

Simbol	Keterangan	Satuan
α	Taraf signifikansi (tingkat kesalahan) pada uji statistik, ditetapkan sebesar 0,05	—
Δ	Delta; selisih nilai metrik varian kuantisasi terhadap <i>baseline</i> FP16	poin / %
H_0	Hipotesis nol (tidak terdapat pengaruh signifikan)	—
H_1	Hipotesis alternatif (terdapat pengaruh signifikan)	—
H_{1a} – H_{1f}	Penanda hipotesis khusus penelitian	—
p	Nilai probabilitas (<i>p-value</i>) hasil uji signifikansi	—
n	Jumlah sampel pengujian per <i>benchmark</i>	sampel
$\pm$	Plus-minus; rentang margin galat (<i>margin of error</i>)	poin
%	Persentase (reduksi memori, retensi akurasi, efisiensi)	persen
$\times$	Operator perkalian	—
$\downarrow$	Penurunan / reduksi nilai	—
$<$	Lebih kecil dari	—
$>$	Lebih besar dari	—
$\leq$	Lebih kecil dari atau sama dengan	—
$\geq$	Lebih besar dari atau sama dengan	—
$^{\circ}\text{C}$	Derajat Celsius (suhu ruangan protokol pengujian)	derajat
RAM_FP16	Konsumsi memori puncak model presisi FP16 (<i>baseline</i>)	MB / GB
RAM_Kuantisasi	Konsumsi memori puncak model setelah kuantisasi	MB / GB
Gen TPS	Kecepatan pembangkitan token (<i>Generation Tokens per Second</i>) pada formula <i>Compound Score</i>	token/detik
-c	Parameter ukuran konteks (<i>context size</i>) pada pengujian Perplexity	token
t/s	Token per detik (satuan kecepatan inferensi)	token/detik
GB	Gigabyte (satuan kapasitas memori/penyimpanan)	gigabyte
MB	Megabyte (satuan kapasitas memori/penyimpanan)	megabyte

DAFTAR GAMBAR

Gambar III. 1	Arsitektur workflow benchmarking SLM/LLM on-device.....	54
Gambar IV. 1	Reduksi ukuran berkas model per varian kuantisasi.	63
Gambar IV. 2	Konsumsi Peak RAM proses per varian kuantisasi pada Helio G99 (RAM 8 GB).	65
Gambar IV. 3	Kecepatan inferensi (Prompt vs Generation) per varian kuantisasi. ..	66
Gambar IV. 4	Degradasi Perplexity pada dataset WikiText-2.....	67
Gambar IV. 5	Hasil benchmark akurasi (MMLU, GSM8K, HumanEval,MT-Bench) untuk kedua model.	69
Gambar IV. 6	Diagram trade-off multi-dimensi (Compound Score) LFM 2.5 dan Qwen 3.5 per varian kuantisasi	76



DAFTAR TABEL

Tabel II. 1 Matriks Perbandingan Penelitian Terkait.....	46
Tabel III. 1 Spesifikasi Perangkat Keras (Hardware).....	55
Tabel III. 2 Spesifikasi Perangkat Lunak (Software)	55
Tabel IV. 1 Reduksi Ukuran Berkas Model per Varian Kuantisasi	62
Tabel IV. 2 Performa Inferensi (Total Waktu, Prompt Speed, Generation Speed, Peak RAM proses, dan CPU Peak).....	64
Tabel IV. 3 Hasil Perplexity (PPL) WikiText-2 per Varian Kuantisasi	67
Tabel IV. 4 Hasil Benchmark Akurasi LFM 2.5 (1,2B)	69
Tabel IV. 5 Hasil Benchmark Akurasi Qwen 3.5 (2B).....	69
Tabel IV. 6 Selisih (Delta) Akurasi LFM 2.5 per Varian terhadap Baseline FP16 (basis Tabel IV.4).....	77
Tabel IV. 7 Selisih (Delta) Akurasi Qwen 3.5 per Varian terhadap Baseline FP16 ..	78
Tabel IV. 8 Matriks Rekomendasi Varian LFM 2.5 untuk Tecno Pova 5 (RAM 8 GB, Helio G99).....	79
Tabel IV. 9 Matriks Rekomendasi Varian Qwen 3.5 untuk Tecno Pova 5 (RAM 8 GB, Helio G99)	80
Tabel IV. 10 Ringkasan Uji Beda Performa Antar Varian LFM 2.5.....	82
Tabel IV. 11 Ringkasan Uji Beda Performa Antar Varian Qwen 3.5	83

DAFTAR LAMPIRAN

LAMPIRAN A SKRIP OTOMATISASI PENGUJIAN (RTX 3060)	100
LAMPIRAN B SKRIP EKSEKUSI INFERENSI EDGE (Tecno Pova 5)	132
LAMPIRAN C HASIL BENCHMARK	136
LAMPIRAN D LANGKAH-LANGKAH PERSIAPAN LINGKUNGAN DAN KUANTISASI MODEL	140
LAMPIRAN E DAFTAR AKRONIM DAN ISTILAH TEKNIS	144
LAMPIRAN F SCREENSHOT TERMINAL	146

