

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan serangkaian eksperimen dan analisis data mengenai optimasi *Post-Training Quantization* (PTQ) pada *Small Language Model* (SLM) untuk implementasi *Edge Computing* berbasis Android, ditarik kesimpulan secara sistematis untuk menjawab kelima butir perumusan masalah penelitian sebagai berikut:

1. Performa Awal Baseline FP16 pada Platform Edge Android

Performa awal (*baseline*) LFM 2.5 (1,2B) dan Qwen 3.5 (2B) dalam format presisi murni *Floating-Point 16-bit* (FP16) pada platform *Edge Computing* berbasis Android menunjukkan konsumsi sumber daya yang sangat berat dan rentan mengalami kegagalan sistem. Model LFM 2.5 FP16 dengan ukuran berkas 2,2 GB mengonsumsi *peak* RAM sebesar 2.303,47 MB (2,30 GB) dan menghasilkan *generation speed* 5,57 TPS. Sementara itu, model Qwen 3.5 FP16 dengan ukuran berkas 3,6 GB mengonsumsi *peak* RAM mencapai 3.744,25 MB (3,74 GB) atau menyerap 46,8% dari kapasitas RAM sistem 8 GB, dengan *generation speed* 1,87 TPS. Tingginya beban alokasi memori pada format FP16 ini menempatkan gawai pada risiko ekstrem memicu intervensi *Out of Memory* (OOM) *Killer* dan penghentian proses secara paksa (*Force Close*).

2. Pengaruh PTQ Terhadap Ukuran Model dan Penggunaan RAM

Penerapan *Post-Training Quantization* (PTQ) berformat GGUF *k-quants* (Q3_K_M, Q4_K_M, Q5_K_M) terbukti secara signifikan ($p\text{-value} < 0,05$) mereduksi dimensi fisik berkas model sebesar 61,1% hingga 74,0% (ukuran berkas menyusut dari 3,6 GB menjadi 1,1–1,4 GB pada Qwen 3.5 dan dari 2,2 GB menjadi 573–805 MB pada LFM 2.5). Pada alokasi memori utama, PTQ secara nyata mengefisienkan penggunaan *peak* RAM proses sebesar 27,7% hingga 60,4% pada LFM 2.5 (turun dari 2.303,47 MB menjadi 911,97–1.665,28 MB) serta sebesar 23,7% hingga 49,3% pada Qwen 3.5 (turun dari 3.744,25 MB menjadi 1.897,97–2.856,29 MB). Penurunan konsumsi RAM secara masif ini terbukti efektif mengamankan memori gawai dan mengeliminasi risiko kegagalan sistem akibat *Out of Memory*.

3. Peningkatan Kecepatan Inferensi (Prompt Processing dan Generation TPS)

Penerapan PTQ terbukti mengakselerasi kecepatan pembangkitan teks (*generation speed*) secara signifikan pada CPU ARM MediaTek Helio G99, yakni meningkat hingga 2,45 kali lipat pada LFM 2.5 Q4_K_M (dari 5,57 TPS menjadi 13,67 TPS) dan 2,65 kali lipat pada Qwen 3.5 Q4_K_M (dari 1,87 TPS menjadi 4,95 TPS). Namun demikian, model Qwen 3.5 mengalami gejala saturasi *throughput* pada kisaran 4–5 TPS akibat limitasi skala arsitektur 2B pada CPU seluler. Selain itu, ditemukan anomali pada *prompt speed*, di mana varian kompresi bit ganjil Q3_K_M justru tercatat paling lambat pada kedua model (anjlok ke 18,27 TPS pada LFM dan 14,55 TPS pada Qwen, berbanding jauh dengan Q4_K_M sebesar 43,67 TPS pada LFM dan 27,00 TPS pada Qwen). Anomali ini membuktikan bahwa kuantisasi bit ganjil menambah beban komputasi *dequantization* pada register CPU ARM, yang menggeser

bottleneck dari *memory-bound* menjadi *compute-bound* serta meningkatkan beban *CPU peak* hingga 464% (LFM) dan 525% (Qwen).

4. Dampak Terhadap Kualitas Linguistik dan Retensi Akurasi Kognitif

Dampak kompresi terhadap kualitas linguistik dan retensi akurasi kognitif menunjukkan tingkat sensitivitas yang bervariasi bergantung pada presisi bit. Pada varian Q4_K_M, degradasi *Perplexity* (PPL) pada dataset WikiText-2 terpantau sangat minor dengan kenaikan 0,53 poin pada LFM 2.5 (dari 12,68 menjadi 13,21) dan 0,59 poin pada Qwen 3.5 (dari 12,78 menjadi 13,36), yang keduanya berada di bawah ambang batas toleransi 1 poin. Keragaman leksikal (MT-Bench TTR) mengalami penyempitan moderat pada Q4_K_M dari 0,526 menjadi 0,468 (LFM) dan dari 0,531 menjadi 0,523 (Qwen). Pada retensi akurasi kognitif rata-rata (MMLU, GSM8K, HumanEval), Q4_K_M mampu mempertahankan retensi sebesar 88,9% pada LFM 2.5 (skor 112/300 sampel vs. *baseline* 126/300) dan 100,0% sempurna pada Qwen 3.5 (skor 119/300 sampel vs. *baseline* 119/300, didukung lonjakan akurasi GSM8K dari 27% ke 42%). Sebaliknya, kuantisasi ekstrem Q3_K_M memicu degradasi kognitif fatal pada tugas eksak, ditandai anjloknya akurasi matematis GSM8K sebesar 18 poin pada LFM 2.5 (dari 58% menjadi 40%) serta anjloknya akurasi pemrograman HumanEval sebesar 27 poin pada Qwen 3.5 (dari 52% menjadi 25%).

5. Penetapan Konfigurasi dan Tingkat Kompresi Optimal (Sweet Spot)

Berdasarkan analisis pertukaran multi-dimensi (*multi-dimensional trade-off*), varian **Q4_K_M** secara konsisten terbukti sebagai konfigurasi tingkat kompresi paling optimal (*sweet spot*) untuk diimplementasikan secara murni dan stabil pada perangkat Android berkapasitas RAM 8 GB ditenagai SoC

MediaTek Helio G99. Varian Q4_K_M menyajikan titik kesetimbangan terbaik lintas-arsitektur dengan mereduksi ukuran berkas sebesar 66%–68%, mengefisienkan *peak* RAM sebesar 31,6%–36,9%, mengakselerasi *generation speed* hingga 2,45–2,65 kali lipat, serta menjaga retensi nalar kognitif pada tingkat yang sangat tinggi (88,9% pada LFM 2.5 dan 100,0% pada Qwen 3.5). Varian Q4_K_M secara sempurna menghindari kelemahan varian Q3_K_M yang mengalami anomali latensi *prompt speed*, beban puncak CPU ekstrem yang memicu *thermal throttling*, serta kerusakan fatal pada logika matematika dan kode pemrograman.

Hasil analisis data secara menyeluruh menunjukkan bahwa keenam hipotesis khusus yang diajukan dalam penelitian ini terbukti kebenarannya secara empiris, dengan satu temuan yang arah pengaruhnya berbeda dari proyeksi awal. Penerapan PTQ terbukti mengurangi ukuran berkas (H1a), menurunkan *peak* RAM (H1b), serta mempercepat *Generation TPS* (H1d) secara signifikan pada kedua arsitektur. Varian Q4_K_M terkonfirmasi sebagai titik *sweet spot* operasional dengan retensi kapabilitas kognitif yang sangat tinggi, yakni 88,9% pada LFM 2.5 dan 100,0% pada Qwen 3.5, sekaligus menjaga degradasi *Perplexity* tetap di bawah ambang toleransi 1 poin (kenaikan 0,53 poin pada LFM 2.5 dan 0,59 poin pada Qwen 3.5); hal ini memvalidasi H1e dan H1f. Sebaliknya, varian kompresi tertinggi (Q3_K_M) memicu degradasi fungsional yang signifikan pada tugas kognitif eksak, ditandai penurunan GSM8K sebesar 18 poin pada LFM 2.5 dan anjloknya HumanEval sebesar 27 poin pada Qwen 3.5. Adapun H1c terbukti dalam arti terdapat pengaruh signifikan PTQ terhadap utilitas CPU, namun arah pengaruhnya berupa peningkatan beban CPU peak (hingga 464% pada LFM 2.5 dan 525% pada Qwen 3.5), bukan penurunan sebagaimana

diproyeksikan; hal ini disebabkan oleh pergeseran *bottleneck* dari *memory-bound* menjadi *compute-bound*. Dengan demikian, penetapan varian Q4_K_M sebagai titik keseimbangan (*trade-off*) paling optimal tetap tervalidasi secara kuat oleh bukti empiris pada kedua arsitektur model uji.

5.2. Saran

Berdasarkan temuan penelitian, hasil uji statistik, serta keterbatasan yang teridentifikasi sepanjang eksperimen, penulis mengajukan rekomendasi pengembangan yang dikelompokkan ke dalam tiga aspek sebagaimana panduan penulisan, yaitu aspek manajerial, aspek sistem, dan aspek penelitian lanjutan. Setiap rekomendasi dipasangkan dengan keterbatasan yang relevan agar dapat ditelusuri konteksnya.

5.2.1. Aspek Manajerial

1. Adopsi Q4_K_M sebagai Konfigurasi Produksi. Bagi para pengembang aplikasi *Mobile AI* yang menargetkan ekosistem perangkat Android kelas menengah dengan kapasitas RAM 8 GB, varian Q4_K_M sangat direkomendasikan untuk diadopsi sebagai konfigurasi baku (*default*). Berdasarkan data empiris pada Tabel IV.8, varian kompresi ini terbukti mampu mereduksi penggunaan *peak RAM* proses sebesar 36,9% sekaligus mengakselerasi *generation speed* hingga 2,45 kali lipat dibandingkan dengan *baseline* FP16. Keunggulan efisiensi ini juga dibarengi dengan stabilitas fungsional yang aman, di mana tingkat retensi akurasi rata-rata model mampu bertahan di angka 88,9% dengan beban *CPU peak* yang tetap terkontrol di kisaran 300%. Nilai beban komputasi tersebut dinilai masih berada jauh di bawah ambang batas kritis yang dapat memicu gejala *thermal throttling*,

sehingga mampu menjamin stabilitas performa aplikasi pada perangkat gawai dalam jangka panjang.

2. Pemisahan Tahap Persiapan Model dan Eksekusi *Runtime*. Hasil penelitian menunjukkan bahwa tahap kuantisasi berkas .gguf lebih efisien dilakukan di PC ber-GPU, sedangkan eksekusi *runtime* dilakukan di perangkat *target*. Pola kerja dua-*host* ini disarankan diadopsi sebagai standar pengembangan agar siklus iterasi tetap cepat tanpa mengorbankan kemampuan *deployment on-device*.
3. Penyiapan SOP Pengujian Multi-perangkat. Mengingat hasil penelitian saat ini hanya tervalidasi pada satu unit Tecno Pova 5 (Helio G99), tim pengembang produksi disarankan menyusun *Standard Operating Procedure* pengujian lintas-SoC (Snapdragon 6/7/8 gen, Dimensity, Tensor) sebelum merilis aplikasi berbasis SLM ke pasar.

5.2.2. Aspek Sistem

1. Eksplorasi Akselerator Komputasi Heterogen. Mengingat seluruh inferensi pada penelitian ini bersifat *CPU-bound*, pengembangan selanjutnya disarankan mengintegrasikan kerangka inferensi yang mendukung delegasi beban kalkulasi pada modul AI khusus, seperti *Neural Processing Unit* (NPU) atau akselerasi GPU mobile via Vulkan/OpenCL, agar batas atas TPS dapat didorong lebih jauh.
2. Penambahan Instrumen Telemetry Termal dan Energi. Penelitian belum menyertakan pengukuran suhu CPU dan konsumsi baterai per 1.000 *token* yang diproduksi. Penambahan kedua metrik ini sangat penting untuk evaluasi praktis dalam konteks *Mobile Edge AI* yang sensitif terhadap *thermal throttling* dan ketahanan baterai. Pendekatan pengukuran energi *high-resolution* berbasis

sensor arus eksternal yang dipaparkan oleh (Husom et al., 2025) pada *Raspberry Pi* dapat menjadi acuan adopsi pada perangkat Android.

3. Integrasi Antarmuka Pengguna Grafis (*Native GUI*). Untuk meningkatkan utilitas bagi pengguna akhir, purwarupa terminal ini disarankan dikembangkan ke dalam bentuk aplikasi Android *native* menggunakan *Java Native Interface* (JNI). Tujuannya adalah mentransformasi sistem berbasis CLI menjadi asisten AI interaktif yang lebih intuitif dan aksesibel.

5.2.3. Aspek Penelitian

1. Perluasan Lingkup Perangkat Uji. Replikasi eksperimen pada minimal tiga SoC berbeda (Snapdragon, Dimensity, Tensor) sangat dianjurkan untuk menguji generalisasi temuan *sweet spot* Q4_K_M. Variasi karakteristik termal, *governor* CPU, dan skema *power management* berpotensi menggeser titik keseimbangan optimal antar SoC.
2. Peningkatan Ukuran Sampel *Benchmark* Akurasi. Evaluasi MMLU, GSM8K, dan HumanEval saat ini menggunakan 100 sampel acak per *benchmark*. Studi lanjutan disarankan menggunakan minimal 500 sampel per *benchmark* atau melaporkan interval kepercayaan secara eksplisit agar perbedaan antar varian dapat dipertegas.
3. Komparasi dengan Algoritma Kuantisasi Berbasis Aktivasi. Guna mempertahankan kualitas kognitif pada tingkat kompresi rendah, disarankan melakukan komparasi metode *k-quants* dengan algoritma mutakhir seperti AWQ (Lin et al., 2026) atau MobileQuant (Tan et al., 2024) yang lebih adaptif dalam melindungi bobot *outliers*. Pengembangan lanjutan terhadap teknik *prompting*, *alignment*, dan *inference* yang efisien sebagaimana diuraikan oleh

(Xiao & Zhu, 2025) juga dapat ditambahkan untuk meningkatkan kualitas respons asisten.

