

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Lingkungan Pengujian dan Skenario

Bab ini menguraikan data hasil pengujian (*benchmarking*) yang diperoleh melalui eksekusi model AI sesuai skenario yang telah ditetapkan. Penelitian ini menerapkan metode *Host-to-Target Deployment*: fase kompresi bobot arsitektur melalui PTQ dilakukan pada mesin *host* berspesifikasi tinggi (PC dengan GPU NVIDIA RTX 3060) guna mengoptimalkan kecepatan kalkulasi *K-Quants*. Berkas hasil kompresi dalam format .gguf selanjutnya dipindahkan ke unit *Target Edge Device*, yaitu *smartphone* Android Tecno Pova 5 (SoC MediaTek Helio G99, RAM 8 GB), untuk dieksekusi secara lokal melalui *llama.cpp* pada terminal Termux.

Instrumen komparasi silang (*cross-validation*) menggunakan dua varian SLM, yaitu LFM 2.5 berukuran 1,2 miliar parameter sebagai representasi arsitektur ultra-ringan dan Qwen 3.5 berukuran 2 miliar parameter sebagai instrumen uji beban memori (*stress-test*). Evaluasi dilakukan secara bertingkat dengan membandingkan versi FP16 sebagai variabel kontrol terhadap tiga varian kuantisasi, yakni Q5_K_M, Q4_K_M, dan Q3_K_M.

4.2. Hasil Uji Efisiensi Infrastruktur (Hardware)

Pengujian efisiensi infrastruktur ditujukan untuk mengidentifikasi dampak kuantisasi terhadap reduksi kapasitas penyimpanan fisik, alokasi memori sistem Android (RAM), serta akselerasi kinerja prosesor (TPS).

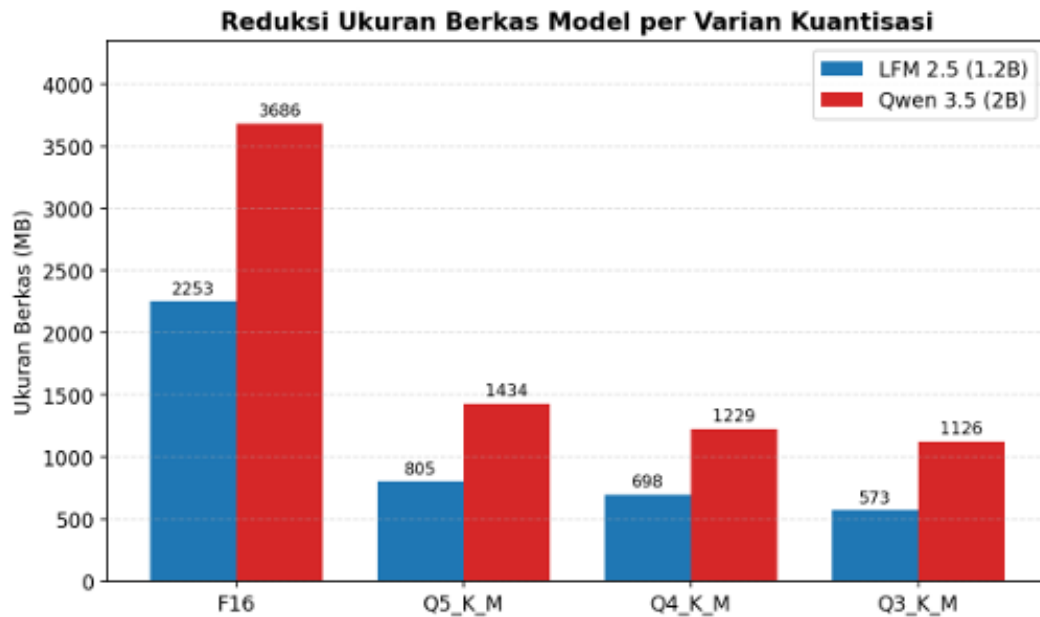
4.2.1. Reduksi Kapasitas Penyimpanan Fisik (*Storage*)

Tahap observasi awal berfokus pada pengukuran ukuran aktual berkas model pada penyimpanan internal gawai. Data komparasi ukuran fisik dan persentase reduksinya disajikan pada Tabel IV.1.

Tabel IV. 1
Reduksi Ukuran Berkas Model per Varian Kuantisasi

Model Arsitektur	Presisi / Format	Ukuran Berkas	Persentase Reduksi
LFM 2.5 (1,2B)	FP16 (Original)	2,2 GB	-
LFM 2.5 (1,2B)	Q5_K_M	805 MB	↓ 63,41%
LFM 2.5 (1,2B)	Q4_K_M	698 MB	↓ 68,27%
LFM 2.5 (1,2B)	Q3_K_M	573 MB	↓ 73,95%
Qwen 3.5 (2B)	FP16 (Original)	3,6 GB	-
Qwen 3.5 (2B)	Q5_K_M	1,4 GB	↓ 61,11%
Qwen 3.5 (2B)	Q4_K_M	1,2 GB	↓ 66,67%
Qwen 3.5 (2B)	Q3_K_M	1,1 GB	↓ 69,44%

Sumber: (Penulis)



Sumber: (Penulis)

Gambar IV. 1
Reduksi ukuran berkas model per varian kuantisasi.

Berdasarkan Tabel IV.1, implementasi kuantisasi mampu mereduksi ukuran berkas model secara signifikan melampaui 60%. Pada varian Q3_K_M, ukuran LFM 2.5 berhasil ditekan hingga 573 MB, sementara Qwen 3.5 menyusut menjadi 1,1 GB, sehingga memberikan ketersediaan ruang penyimpanan ROM yang lebih besar pada perangkat.

4.2.2. Konsumsi RAM dan Kecepatan Inferensi (Helio G99)

Pengujian performa komputasi dilakukan secara *batch* langsung pada perangkat target Tecno Pova 5 melalui *script* otomatisasi *benchmark.sh* di Termux. Pengujian pada tahap ini berfokus murni pada efisiensi *runtime* fisik gawai, terisolasi secara total dari evaluasi skor akurasi kognitif yang telah didelegasikan pada perangkat *host* PC. Setiap varian model dieksekusi sebanyak tiga kali. Guna mengonfrontasi keterbatasan ketiadaan sensor telemetri suhu internal pada lingkungan *non-root*

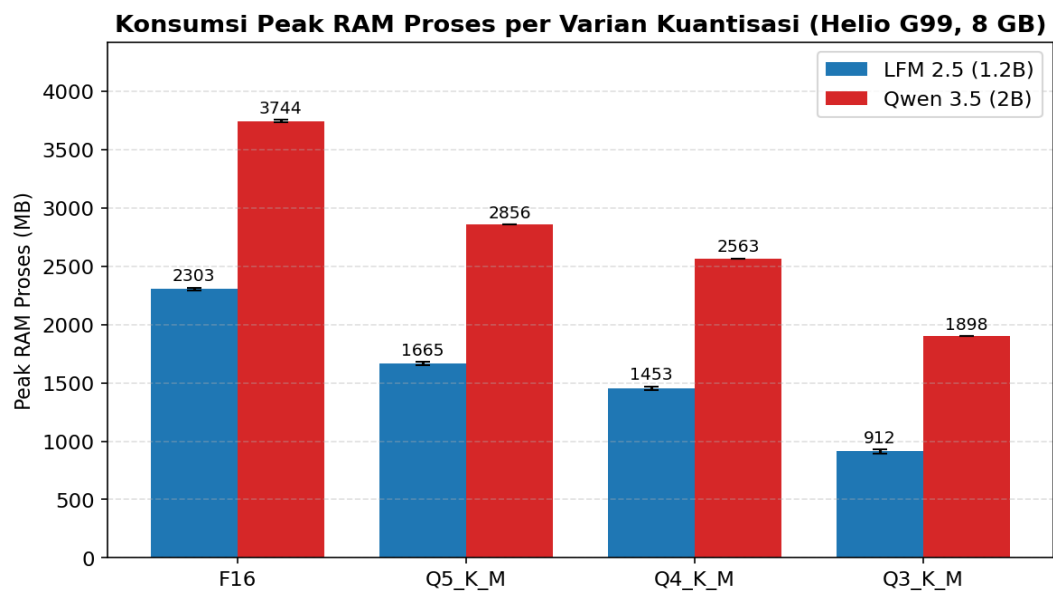
Termux, penelitian ini menyelenggarakan Protokol Ekualisasi Termal (*Thermal Equilibrium Protocol*) berupa jeda *cooldown* manual selama 3–5 menit antar-sesi eksekusi dengan suhu ruangan 24°C. Protokol ini esensial untuk mengembalikan kondisi termal SoC MediaTek Helio G99 ke titik setimbang demi mencegah intervensi *thermal throttling* kernel Android, sehingga fluktuasi nilai TPS murni merepresentasikan dampak deterministik dari tingkat kompresi model. Metrik *Prompt Speed* dan *Generation Speed* diekstraksi dari keluaran statistik llama-cli, sedangkan konsumsi RAM proses (VmRSS) dan beban CPU dipantau secara periodik (setiap 0,5 detik) dari `/proc/<pid>/status` menggunakan konfigurasi 6 *threads*. Hasil pengukuran kuantitatif disajikan pada Tabel IV.2.

Tabel IV. 2
Performa Inferensi (Total Waktu, Prompt Speed, Generation Speed, Peak RAM proses, dan CPU Peak)

Model (Format)	N	Total Waktu (s)	Prompt Speed (t/s)	Gen Speed (t/s)	Peak RAM Proses (MB)	CPU Peak (%)
LFM 2.5 (FP16)	3	16,00	33,47	5,57	2.303,47	292,33
LFM 2.5 (Q5_K_M)	3	10,67	35,33	10,63	1.665,28	357,33
LFM 2.5 (Q4_K_M)	3	7,33	43,67	13,67	1.452,97	299,33
LFM 2.5 (Q3_K_M)	3	14,33	18,27	11,07	911,97	464,00
Qwen 3.5 (FP16)	3	310,33	22,70	1,87	3.744,25	430,33
Qwen 3.5 (Q5_K_M)	3	181,00	22,55	4,60	2.856,29	509,00
Qwen 3.5 (Q4_K_M)	3	297,00	27,00	4,95	2.562,88	500,50

Model (Format)	N	Total Waktu (s)	Prompt Speed (t/s)	Gen Speed (t/s)	Peak RAM Proses (MB)	CPU Peak (%)
Qwen 3.5 (Q3_K_M)	3	187,00	14,55	4,25	1.897,97	525,00

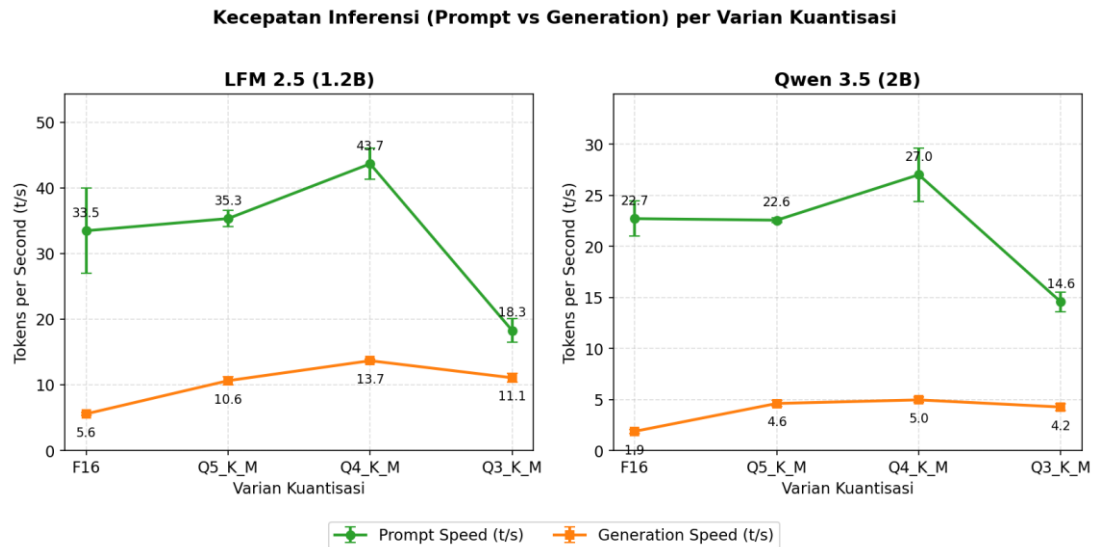
Sumber: (Penulis)



Sumber: (Penulis)

Gambar IV. 2

Konsumsi Peak RAM proses per varian kuantisasi pada Helio G99 (RAM 8 GB).



Sumber: (Penulis)

Gambar IV. 3
Kecepatan inferensi (Prompt vs Generation) per varian kuantisasi.

Tabel IV.2 memperlihatkan tiga pola utama. Pertama, reduksi RAM konsisten pada kedua model: Qwen 3.5 FP16 menyerap 3,74 GB RAM proses atau setara 47% RAM sistem 8 GB dan rawan memicu *OOM Killer*, sedangkan kuantisasi menurunkannya menjadi 2,56 GB pada Q4_K_M (efisiensi 31,6%) dan 1,90 GB pada Q3_K_M (efisiensi 49,3%). Kedua, akselerasi *Generation Speed* dominan pada LFM 2.5: varian Q4_K_M melonjak ke 13,67 t/s atau 2,45 kali lipat *baseline* FP16 (5,57 t/s), sekaligus mencatat total waktu eksekusi tercepat (7,33 detik) dengan CPU *peak* yang terkontrol di kisaran 299%. Ketiga, Qwen 3.5 jenuh di rentang 1,87–4,95 t/s, profil yang mencerminkan keterbatasan model *reasoning* berukuran dua miliar parameter pada CPU ARM kelas menengah ke bawah.

4.3. Hasil Uji Degradasi Kognitif (Software / AI)

Evaluasi kognitif bertujuan mengukur dampak kompresi terhadap kecerdasan *neural network model* menggunakan metrik *Perplexity* serta serangkaian uji pemahaman akademik.

4.3.1. Evaluasi *Perplexity* (PPL)

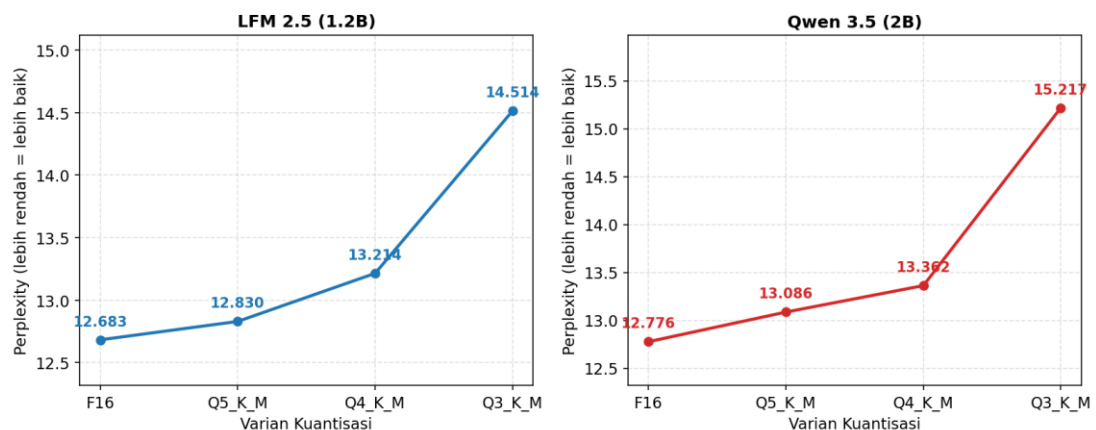
Perplexity digunakan untuk mengukur tingkat ambiguitas model terhadap struktur sintaksis pada *dataset* WikiText-2. Nilai PPL yang lebih rendah mengindikasikan tingkat pemahaman bahasa yang lebih baik. Hasilnya ditunjukkan pada Tabel IV.3.

Tabel IV. 3
Hasil *Perplexity* (PPL) WikiText-2 per Varian Kuantisasi

Varian Kompresi	Ukuran Konteks (-c)	PPL LFM 2.5 (1,2B)	PPL Qwen 3.5 (2B)
FP16 (<i>baseline</i>)	512 Token	12,6829	12,7763
Q5 K M	512 Token	12,8297	13,0856
Q4 K M	512 Token	13,2145	13,3617
Q3 K M	512 Token	14,5135	15,2172

Sumber: (Penulis)

Degradasi *Perplexity* pada Dataset WikiText-2



Sumber: (Penulis)

Gambar IV. 4
Degradasi *Perplexity* pada dataset WikiText-2.

4.3.2. Evaluasi Akurasi Logika (MMLU, GSM8K, dan HumanEval) serta Kreativitas Bahasa (MT-Bench TTR)

Evaluasi akurasi menggunakan tiga instrumen *benchmark* kognitif: MMLU (pemahaman umum berbasis pilihan ganda A/B/C/D), GSM8K (nalar matematika dengan jawaban numerik), dan HumanEval (akurasi pemrograman Python dievaluasi melalui *unit-test*), serta satu instrumen tambahan berupa MT-Bench dengan skor *Type-Token Ratio* (TTR) yang mengukur keragaman leksikal teks generatif. Masing-masing *benchmark* mengeksekusi 100 sampel acak dengan suhu *sampling* deterministik (*temperature=0*). Guna memastikan validitas dan stabilitas pengukuran, eksekusi 100 sampel pada setiap instrumen tersebut diulang sebanyak tiga kali (3 *run*) untuk masing-masing varian model.

Eksekusi pengujian akurasi dijalankan pada perangkat *host* PC NVIDIA RTX 3060 guna mengoptimalkan waktu evaluasi, berlandaskan pada asumsi ekuivalensi statistik. Kendati terdapat perbedaan arsitektur pemrosesan matriks *floating-point* antara arsitektur CUDA (GPU) dan ARM (CPU) yang berpotensi memicu deviasi pembulatan *token* berskala minor, penggunaan format berkas *.gguf* yang seragam menjamin bahwa keluaran persentase akurasi kognitif secara makroskopis tetap berada di dalam ambang batas *margin of error* yang dapat ditoleransi saat model dieksekusi pada perangkat target. Data hasil pengujian disajikan pada Tabel IV.4 dan Tabel IV.5. Angka yang tercantum di dalam kedua tabel tersebut merupakan nilai rerata (*mean*) yang diperoleh dari tiga kali perulangan pengujian (3 *run*) pada setiap varian model guna memastikan stabilitas performa.

Tabel IV. 4
Hasil Benchmark Akurasi LFM 2.5 (1,2B)

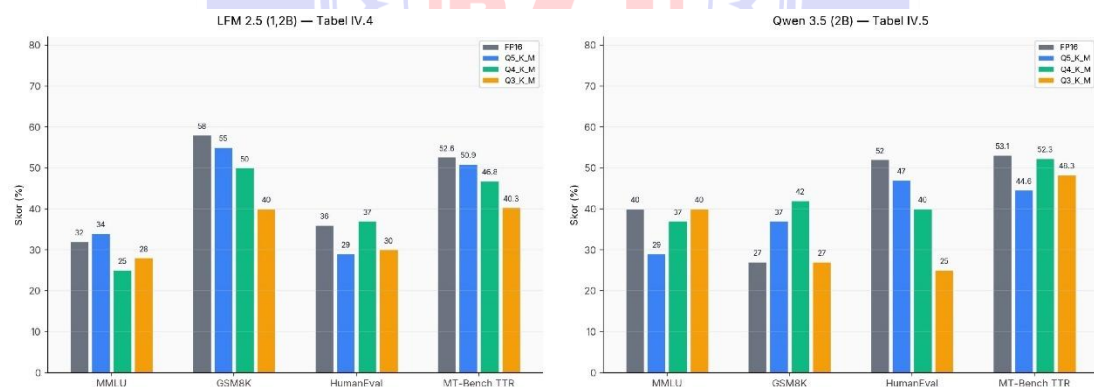
Benchmark	FP16	Q5_K_M	Q4_K_M	Q3_K_M
MMLU (A/B/C/D)	32%	34%	25%	28%
GSM8K (numerik)	58%	55%	50%	40%
HumanEval (kode)	36%	29%	37%	30%
MT-Bench TTR (Type-Token Ratio)	0,526	0,509	0,468	0,403

Sumber: (Penulis)

Tabel IV. 5
Hasil Benchmark Akurasi Qwen 3.5 (2B)

Benchmark	FP16	Q5_K_M	Q4_K_M	Q3_K_M
MMLU (A/B/C/D)	40%	29%	37%	40%
GSM8K (numerik)	27%	37%	42%	27%
HumanEval (kode)	52%	47%	40%	25%
MT-Bench TTR (Type-Token Ratio)	0,531	0,446	0,523	0,483

Sumber: (Penulis)



Sumber: (Penulis)

Gambar IV. 5
Hasil benchmark akurasi (MMLU, GSM8K, HumanEval, MT-Bench) untuk kedua model.

Berdasarkan data pada Tabel IV.4, performa model LFM 2.5 tergolong relatif stabil pada mayoritas pengujian. Namun, metrik nalar matematis (GSM8K) terbukti paling sensitif terhadap kompresi, ditandai dengan penurunan akurasi dari 58% pada presisi FP16 menjadi 40% pada varian Q3_K_M. Sementara itu, Tabel IV.5

memperlihatkan pola yang berbeda pada arsitektur Qwen 3.5. Varian Q4_K_M justru mencatatkan anomali positif dengan skor GSM8K tertinggi (42%) dibandingkan dengan *baseline* FP16 (27%), meskipun akurasi logika pemrograman (HumanEval) pada varian tersebut mengalami penurunan dari 52% menjadi 40%. Lebih lanjut, varian Q3_K_M menunjukkan degradasi performa yang paling signifikan; akurasi HumanEval pada tingkat kompresi ini anjlok tajam menjadi 25%, terlepas dari skor MMLU dan GSM8K yang masih mampu bertahan setara dengan *baseline*.

4.3.3. Karakteristik Hasil pada *Reasoning Model* (Qwen 3.5)

Hasil evaluasi model Qwen 3.5 pada Tabel IV.5 menunjukkan bahwa ketiga *benchmark* utama, yaitu MMLU, GSM8K, dan HumanEval, terbukti stabil pada tiga kali perulangan pengujian (3 *run*). Stabilitas performa ini sejalan dengan karakteristik arsitekturnya, merujuk pada laporan teknis Qwen3 yang memperkenalkan *thinking mode* sebagai mekanisme bawaan keluarga model tersebut (Team, 2025). Skor yang konsisten ini memungkinkan data keluaran dari model Qwen digunakan secara langsung di dalam analisis komparatif tanpa memerlukan perlakuan statistik tambahan.

1. Temuan terpenting pada pengujian model Qwen 3.5 adalah varian Q4_K_M tidak menurunkan rerata akurasi dari ketiga *benchmark* utama. Meskipun skor MMLU mengalami penurunan tipis dari 40% menjadi 37% dan HumanEval turun dari 52% menjadi 40%, skor GSM8K justru melonjak dari 27% menjadi 42%. Kenaikan signifikan pada metrik GSM8K ini membuat rerata akurasi Q4_K_M tetap bertahan kokoh di angka 100,0% jika dikomparasikan terhadap *baseline* FP16.

2. Sebaliknya, varian Q3_K_M dinilai tidak layak untuk dijadikan sebagai varian operasional utama, terlepas dari konsumsi RAM-nya yang paling kecil. Meskipun varian ini mampu mempertahankan skor MMLU di angka 40% dan GSM8K di angka 27%, skor HumanEval justru anjlok 27 poin menjadi 25%. Hal ini mengindikasikan terjadinya kerusakan kognitif yang parah pada kemampuan generasi kode pemrograman.
3. Pada pengukuran keragaman leksikal, metrik MT-Bench TTR model Qwen berfluktuasi dari angka 0,531 pada FP16 menjadi 0,446 pada Q5_K_M, 0,523 pada Q4_K_M, dan 0,483 pada Q3_K_M. Berdasarkan fluktuasi data tersebut, Q4_K_M kembali menjadi varian yang paling mendekati kinerja *baseline* dari sisi keragaman leksikal, sedangkan Q5_K_M justru mencatatkan performa terendah pada metrik TTR.

Secara keseluruhan, data akurasi model Qwen 3.5 menunjukkan bahwa varian Q4_K_M menawarkan *trade-off* yang paling optimal. Varian ini mampu mempertahankan rerata akurasi yang setara dengan presisi murni FP16, bahkan mencatatkan peningkatan pada metrik GSM8K. Selain itu, skor MT-Bench TTR pada varian ini terbukti paling mendekati *baseline*, serta tidak mengalami degradasi ekstrem (*collapse*) pada kemampuan penalaran pemrograman (HumanEval) sebagaimana yang dialami oleh varian Q3_K_M.

4.4. Pembahasan Analisis Komparatif

Sub-bab ini membedah signifikansi data hasil pengujian melalui tinjauan teoritis arsitektur *Edge Intelligence* dan hukum ketahanan kompresi (*scaling laws*).

4.4.1. Efisiensi RAM dan Peningkatan Generation Speed

Hasil pengujian mengonfirmasi bahwa metode PTQ efektif dalam mengatasi kendala *shared-memory* pada perangkat berkapasitas RAM 8 GB. Pemuatan model FP16 pada Qwen 3.5 menyerap hampir 47% kapasitas RAM sistem (3,74 GB) dan berpotensi memicu *OOM Killer* dari sistem operasi Android. Intervensi Q4_K_M berhasil mereduksi penggunaan RAM Qwen menjadi 2,56 GB (efisiensi 31,6%), sedangkan Q3_K_M menekannya menjadi 1,90 GB (efisiensi 49,3%), sehingga stabilitas operasional perangkat terjaga.

Penurunan beban *bandwidth* data dari RAM ke CPU berimplikasi langsung pada peningkatan *generation speed* untuk LFM 2.5, yang melonjak dari 5,57 TPS pada FP16 menjadi 13,67 TPS pada Q4_K_M, atau setara akselerasi 2,45 kali lipat dibanding versi murninya. Total waktu eksekusi LFM Q4_K_M juga menjadi yang tersingkat (7,33 detik) atau 2,18 kali lebih cepat dibanding FP16 (16 detik). Pola ini konsisten dengan dalil *Memory-Bound* (Zhang et al., 2024) sekaligus sejalan dengan studi sistematis pertama eksekusi LLM *on-device* pada Android dan iOS yang menyimpulkan bahwa inferensi LLM bersifat *memory-bound* (Laskaridis et al., 2024). Khusus untuk LFM 2.5, percepatan turut diperkuat oleh karakteristik *hybrid backbone* yang dirancang untuk *fast prefill/decode* di CPU (Amini et al., 2025).

4.4.2. Anomali Kecepatan Baca (*Prompt Speed*) pada Varian Q3_K_M

Secara teoritis, varian dengan konsumsi RAM paling kecil seharusnya menghasilkan kecepatan inferensi tertinggi, namun Tabel IV.2 menunjukkan pola yang justru sebaliknya pada Q3_K_M. Pada LFM 2.5, *Prompt Speed* Q3_K_M anjlok ke 18,27 t/s, jauh di bawah Q4_K_M (43,67 t/s) maupun Q5_K_M (35,33 t/s). Pola

serupa juga teramati pada Qwen 3.5, dengan Q3_K_M hanya mencapai 14,55 t/s dibanding Q4_K_M (27,00 t/s) dan Q5_K_M (22,55 t/s).

Fenomena tersebut dianalisis sebagai konsekuensi arsitektur CPU ARM. Proses *unpacking* bobot 4-bit dan 5-bit relatif efisien karena strukturnya simetris terhadap *register* CPU. Format 3-bit bersifat ganjil sehingga memaksa set instruksi CPU melakukan operasi *bit-shifting* tambahan yang kompleks, sehingga muncul *bottleneck* pada fase *pre-fill* dan memperpanjang durasi eksekusi total. Penjelasan ini selaras dengan studi resmi Arm Inc. yang mengkonfirmasi bahwa proporsi instruksi *multiplication* bermanfaat pada eksekusi LLM terkuantisasi di CPU Arm sangat rendah karena *cycle* CPU banyak terserap oleh *unpacking* susunan bobot dan *dequantization* (Gope et al., 2024).

Beban *CPU peak* varian Q3_K_M pada kedua model tercatat sebagai yang tertinggi (LFM 2.5 mencapai 464% dan Qwen 3.5 mencapai 525%). Lonjakan ini mengindikasikan terjadinya saturasi *thread* yang lebih intensif akibat tingginya beban komputasi pada operasi *dequantization*. Fenomena ini sejalan dengan studi *real-world* mengenai pengujian LLM terkuantisasi pada berbagai *smartphone*, yang menyoroti besarnya pengaruh manajemen *thermal* dan mikroarsitektur SoC (*System on a Chip*) terhadap capaian *sustained throughput* (Çöplü et al., 2023). Temuan ini menegaskan bahwa PTQ tidak menurunkan beban CPU, melainkan justru meningkatkannya. Hal ini terjadi karena kuantisasi menggeser titik kemacetan (*bottleneck*) dari subsistem memori (*memory-bound*) menuju subsistem komputasi (*compute-bound*): inti CPU tidak lagi banyak mengganggu menunggu transfer data dari RAM, sehingga bekerja lebih padat untuk menghasilkan throughput token yang lebih tinggi, ditambah overhead aritmetika dari proses *dequantization* (Gope et al., 2024). Dengan demikian,

hipotesis H1c terbukti terdapat pengaruh signifikan PTQ terhadap utilitas CPU meskipun arah pengaruhnya berupa peningkatan, bukan penurunan.

4.4.3. Dampak Distorsi terhadap Nalar Matematika dan Logika Pemrograman

Selain anomali kecepatan, varian Q3_K_M juga memperlihatkan tanda kerusakan pada struktur nalar model. Nilai *Perplexity* yang melampaui ambang toleransi dikuatkan oleh anjloknya akurasi GSM8K pada LFM 2.5 (dari 58% menjadi 40%, selisih 18 poin) dan HumanEval pada Qwen 3.5 (dari 52% menjadi 25%, selisih 27 poin). Penurunan tajam tersebut disebabkan oleh hilangnya nilai-nilai pencilan (*outliers*) pada matriks bobot akibat kuantisasi ekstrem (Gong et al., 2024). Tidak seperti bahasa naratif yang toleran terhadap redundansi, logika matematis dan pemrograman bersifat eksak, sehingga pemotongan presisi yang terlalu agresif secara langsung merusak logika fungsional model.

Dampak dari kuantisasi ekstrem ini juga terproyeksi secara linear pada penurunan skor MT-Bench TTR. Nilai TTR untuk LFM 2.5 mengalami penurunan dari 0,526 pada presisi murni FP16 menjadi 0,403 pada varian Q3_K_M, sedangkan Qwen 3.5 turun dari 0,531 menjadi 0,483 pada tingkat kompresi yang sama. Fenomena ini membuktikan bahwa pengurangan bit yang terlalu agresif cenderung menyempitkan keragaman leksikal teks generatif, meskipun magnitudo dampaknya bervariasi bergantung pada karakteristik arsitektur masing-masing model.

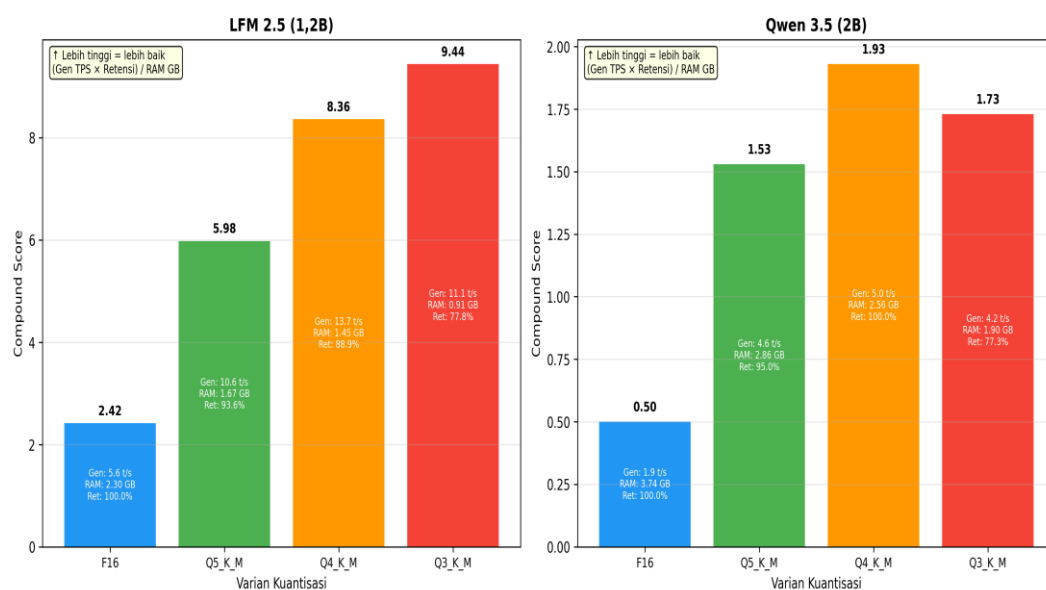
4.4.4. Penetapan Titik Keseimbangan Optimal (*Sweet Spot*)

Berdasarkan analisis komprehensif terhadap akumulasi metrik RAM, TPS, *Perplexity* (PPL), dan akurasi *benchmark*, varian Q4_K_M ditetapkan sebagai titik *sweet spot* optimal untuk implementasi *on-device*. Pada pengujian model LFM 2.5,

varian Q4_K_M mampu menyajikan *generation speed* tertinggi mencapai 13,67 TPS dengan tetap mempertahankan retensi akurasi kognitif di angka 88,9%. Pola keunggulan ini terulang secara konsisten pada arsitektur Qwen 3.5, di mana varian Q4_K_M memberikan *generation speed* tertinggi sebesar 4,95 TPS mampu mempertahankan rerata akurasi kumulatif secara ekuivalen (100,0%) terhadap *baseline* FP16, meskipun terjadi fluktuasi asimetris di mana penurunan pada tugas logika pemrograman berhasil dikompensasi oleh lonjakan performa pada nalar matematis.

Sebaliknya, varian Q3_K_M resmi dieliminasi dari rekomendasi operasional utama. Penolakan terhadap varian kompresi tertinggi ini didasarkan pada performa *prompt speed* yang rendah, lonjakan beban *CPU peak* yang berisiko tinggi memicu *thermal throttling*, serta tingkat degradasi akurasi yang terlalu besar sehingga menurunkan reliabilitas model secara keseluruhan saat mengeksekusi instruksi kognitif.

Diagram Trade-off Multi-dimensi (Compound Score) per Varian Kuantisasi



Sumber: Penulis

Gambar IV. 6
Diagram trade-off multi-dimensi (Compound Score) LFM 2.5 dan Qwen 3.5 per
varian kuantisasi

4.4.5. Kuantifikasi Penurunan Kepintaran FP16 vs Q3/Q4/Q5 dan Rekomendasi Varian Operasional untuk Tecno Pova 5

Sub-bab ini menjawab dua pertanyaan praktis sekaligus, yakni seberapa besar kepintaran model yang hilang akibat kuantisasi dan varian mana yang paling tepat dijalankan pada Tecno Pova 5. Pendekatan analitik dilakukan dengan memadukan dua sudut pandang, yaitu retensi kepintaran (akurasi relatif terhadap *baseline* FP16) dan efisiensi operasional (RAM, kecepatan inferensi, serta margin keamanan termal). Kuantifikasi disajikan secara paralel untuk dua keluarga model uji, yaitu LFM 2.5 (1,2 miliar parameter) dan Qwen 3.5 (2 miliar parameter).

Sesuai dengan ketentuan taksonomi metodologi yang telah dikunci pada Subbab 3.4, analisis kuantifikasi penurunan kapabilitas cerdas ini menerapkan dua regulasi evaluasi yang ketat. Pertama, metrik MT-Bench TTR diposisikan secara eksklusif sebagai indikator keragaman leksikal (*lexical diversity*) dan dievaluasi secara terisolasi di luar kalkulasi rata-rata akurasi tugas kognitif. Kedua, nilai performa kognitif komposit untuk LFM 2.5 dan Qwen 3.5 dihitung secara homogen berdasarkan nilai rerata (*mean*) dari 3 kali perulangan lintasan pengujian (*3 evaluation runs*) pada perangkat *host* RTX 3060. Khusus untuk arsitektur *reasoning* Qwen 3.5, stabilitas nilai komposit antar-perulangan ini juga didukung oleh karakteristik distribusinya yang terbukti stabil berkat keberadaan *thinking mode* bawaan model (Team, 2025). Langkah standarisasi 3 *runs* ini terbukti efektif dalam memangkas bias variansi skor acak individual, sehingga visualisasi dampak deviasi presisi bit (FP16 terhadap varian Q5,

Q4, dan Q3) dapat terkuantisasi secara sah dan dapat dipertanggungjawabkan secara statistik.

Tabel IV.6 menyajikan selisih akurasi LFM 2.5 untuk masing-masing varian kuantisasi terhadap *baseline* FP16. Kolom Δ mencatat poin persentase yang hilang (atau diperoleh) pada setiap *benchmark*, lalu ditutup oleh dua baris ringkasan yang menggambarkan rata-rata akurasi tiga *benchmark* dan persentase retensi kepintaran terhadap FP16.

Tabel IV. 6
Selisih (Delta) Akurasi LFM 2.5 per Varian terhadap *Baseline* FP16 (basis Tabel IV.4)

<i>Benchmark</i>	FP16 (kontrol)	Q5_K _M	Δ Q5	Q4_K _M	Δ Q4	Q3_K _M	Δ Q3
MMLU (A/B/C/D)	32%	34%	+2	25%	-7	28%	-4
GSM8K (numerik)	58%	55%	-3	50%	-8	40%	-18
HumanEval (kode)	36%	29%	-7	37%	+1	30%	-6
MT-Bench TTR ($\times 100$) *	52,6	50,9	-1,7	46,8	-5,8	40,3	-12,3
Rata-rata akurasi tiga <i>benchmark</i> (MMLU, GSM8K, HumanEval)	42,0%	39,3 %	-2,7	37,3 %	-4,7	32,7 %	-9,3
Retensi kepintaran (% terhadap FP16)	100%	-	93,6 %	-	88,9 %	-	77,8 %

Sumber: (Penulis)

Catatan: MT-Bench TTR dilaporkan secara eksklusif sebagai metrik keragaman leksikal independen dan dikeluarkan dari komputasi komposit rata-rata akurasi tugas kognitif, sesuai dengan ketentuan metodologi pada Subbab 3.4.

Tiga pola utama dapat dibaca dari Tabel IV.6. Pertama, GSM8K menjadi *benchmark* paling sensitif terhadap kuantisasi dengan penurunan monoton dan konsisten lintas varian, yaitu Q5 turun 3 poin, Q4 turun 8 poin, dan Q3 turun 18 poin. Pola ini sejalan dengan temuan (Kurtic et al., 2025) bahwa *task* yang menuntut presisi

numerik dan logika berlapis paling rentan terhadap reduksi presisi bobot karena kuantisasi menghilangkan tepi *outlier* yang berperan sebagai *anchor* numerik halus pada bobot atensi. Kedua, MMLU dan HumanEval mengalami fluktuasi non-monoton, ditandai oleh kenaikan 2 poin pada MMLU Q5 dan kenaikan 1 poin pada HumanEval Q4. Anomali seperti turunnya MMLU Q4 (25%) di bawah Q3 (28%) berada di dalam margin *error* statistik 95% CI sebesar ± 10 poin pada $n=100$, sehingga selisih 3 poin tersebut tidak signifikan secara statistik dan tidak dapat ditafsirkan sebagai indikasi bahwa Q3 lebih cerdas daripada Q4. Ketiga, MT-Bench TTR turun monoton dari 52,6 (FP16) menjadi 40,3 (Q3), menandakan penyempitan distribusi leksikal pada lapisan generatif teks bebas sebagai akibat hilangnya presisi bobot pasca-kuantisasi.

Tabel IV.7 menyajikan informasi serupa untuk Qwen 3.5. Struktur kolomnya identik dengan Tabel IV.6 supaya pembaca dapat membandingkan dua arsitektur secara langsung.

Tabel IV. 7
Selisih (Delta) Akurasi Qwen 3.5 per Varian terhadap *Baseline* FP16

<i>Benchmark</i>	FP16 (kontrol)	Q5_K _M	$\Delta Q5$	Q4_K _M	$\Delta Q4$	Q3_K _M	$\Delta Q3$
MMLU (A/B/C/D)	40%	29%	-11	37%	-3	40%	± 0
GSM8K (numerik)	27%	37%	+10	42%	+15	27%	± 0
HumanEval (kode)	52%	47%	-5	40%	-12	25%	-27
MT-Bench TTR ($\times 100$)	53,1	44,6	-8,5	52,3	-0,8	48,3	-4,8
Rata-rata akurasi tiga <i>benchmark</i> (MMLU, GSM8K, HumanEval)	39,7%	37,7 %	-2,0	39,7 %	± 0	30,7 %	-9,0
Retensi kepintaran (% terhadap FP16)	100%	-	95,0 %	-	100, 0%	-	77,3 %

Sumber: (Penulis)

Catatan: MT-Bench TTR dilaporkan secara eksklusif sebagai metrik keragaman leksikal independen dan dikeluarkan dari komputasi komposit rata-rata akurasi tugas kognitif, sesuai dengan ketentuan metodologi pada Subbab 3.4.

Untuk menjawab pertanyaan praktis “varian mana yang terbaik untuk Tecno Pova 5?”, Tabel IV.8 mengintegrasikan retensi kepintaran LFM 2.5 dari Tabel IV.6 dengan tiga metrik operasional dari Tabel IV.2 (RAM, *Generation Speed*, dan total waktu eksekusi). Hasil integrasi tersebut diolah menjadi *Compound Score* yang menggambarkan *throughput* efektif relatif terhadap konsumsi RAM dan retensi akurasi.

Tabel IV. 8
Matriks Rekomendasi Varian LFM 2.5 untuk Tecno Pova 5 (RAM 8 GB, Helio G99)

Varian	<i>Peak</i> RAM (MB)	Gen TPS (t/s)	Total Waktu (s)	Retensi Akurasi	CPU <i>Peak</i> (%)	<i>Compound</i> <i>Score</i>
FP16	2.303	5,57	16,00	100,0%	292	2,42
Q5_K_M	1.665	10,63	10,67	93,6%	357	5,98
Q4_K_M	1.453	13,67	7,33	88,9%	299	8,36
Q3_K_M	912	11,07	14,33	77,8%	464	9,44

Sumber: (Penulis)

Tabel IV.8 menampilkan dua kandidat dengan *Compound Score* tertinggi, yaitu Q4_K_M (8,36) dan Q3_K_M (9,44). Q3_K_M unggul secara aritmetik karena konsumsi RAM-nya paling kecil, tetapi tiga catatan berikut menjadikan varian tersebut tidak direkomendasikan untuk Tecno Pova 5. Pertama, performa kognitif mengalami degradasi drastis pada nalar matematika dengan GSM8K turun 18 poin, sehingga model praktis kehilangan kemampuan menyelesaikan soal hitung sederhana yang merupakan *use case* utama asisten AI di perangkat seluler. Kedua, CPU *peak* mencapai 464% atau setara 77% saturasi enam *thread* yang dialokasikan, sehingga memperbesar

risiko *thermal throttling* pada sesi panjang yang pada *smartphone* Helio G99 tanpa pendingin aktif berakibat penurunan *sustained throughput* (Çöplü et al., 2023). Ketiga, terjadi anomali *Prompt Speed* sebesar 18,27 t/s yang lebih rendah dibanding Q4_K_M (43,67 t/s) akibat *unpacking* susunan bit ganjil yang sudah dianalisis pada Sub-bab 4.4.2 dan didukung oleh (Gope et al., 2024).

Berdasarkan tiga catatan tersebut, Q4_K_M ditetapkan sebagai varian operasional paling tepat untuk LFM 2.5 di Tecno Pova 5. Kepintaran yang hilang pada varian ini bersifat moderat dengan rata-rata penurunan 4,7 poin akurasi atau retensi 88,9% terhadap FP16, namun *trade-off* yang diperoleh sangat menguntungkan, yaitu *peak* RAM turun 36,9%, *Generation Speed* meningkat 2,45 kali lipat, total waktu eksekusi turun 54,2%, dan CPU *peak* tetap terkendali di kisaran 300%. Kombinasi ini memberi margin aman bagi RAM 8 GB Tecno Pova 5 yang masih harus melayani OS Android dan aplikasi latar belakang, sekaligus meminimalkan risiko *thermal throttling* pada SoC Helio G99.

Tabel IV.9 mengintegrasikan retensi kepintaran Qwen 3.5 dari Tabel IV.7 dengan metrik operasional dari Tabel IV.2 menggunakan rumus *Compound Score* yang sama dengan Tabel IV.8.

Tabel IV. 9
Matriks Rekomendasi Varian Qwen 3.5 untuk Tecno Pova 5 (RAM 8 GB, Helio G99)

Varian	<i>Peak</i> RAM (MB)	Gen TPS (t/s)	Total Waktu (s)	Retensi Akurasi	CPU <i>Peak</i> (%)	<i>Compound</i> <i>Score</i>
FP16	3.744	1,87	310,33	100,0%	430	0,50
Q5_K_M	2.856	4,60	181,00	95,0%	509	1,53
Q4_K_M	2.563	4,95	297,00	100,0%	501	1,93
Q3_K_M	1.898	4,25	187,00	77,3%	525	1,73

Sumber: (Penulis)

Berdasarkan data pada Tabel IV.9, varian Q4_K_M terbukti menjadi pilihan terbaik untuk arsitektur Qwen 3.5 dengan perolehan *Compound Score* tertinggi sebesar 1,93. Angka ini mengungguli capaian varian Q5_K_M (1,53) dan varian Q3_K_M (1,73). Sebaliknya, varian Q3_K_M tidak direkomendasikan untuk penggunaan harian karena memicu degradasi kognitif dan efisiensi yang masif, di mana skor HumanEval anjlok sebesar 27 poin, beban *CPU peak* melonjak drastis hingga mencapai 525%, serta *prompt speed* yang tersendat di angka 14,55 TPS. Dengan demikian, keputusan rekomendasi operasional untuk penerapan model Qwen 3.5 pada perangkat gawai pintar Tecno Pova 5 secara mutlak jatuh pada varian Q4_K_M.

Konvergensi rekomendasi pada Q4_K_M memperkuat justifikasi penetapannya sebagai varian *default* lintas-arsitektur. Pada LFM 2.5 yang mengusung arsitektur *hybrid non-reasoning*, Q4_K_M menggabungkan retensi kepintaran 88,9% dengan *Compound Score* 8,36 dan *Generation Speed* 13,67 t/s. Pada Qwen 3.5 yang berarsitektur *reasoning* dua-mode, Q4_K_M mempertahankan retensi 100,0% dengan *Compound Score* 1,93 dan *Generation Speed* 4,95 t/s. Karena rekomendasi konsisten pada kedua keluarga model, Q4_K_M dapat ditetapkan sebagai konfigurasi *default* yang siap diadopsi pengembang tanpa harus melakukan kalibrasi ulang ketika berpindah arsitektur, selama target perangkat tetap Android berbasis Helio G99 dengan RAM 8 GB.

Perlu ditegaskan bahwa penetapan *Compound Score* ini merupakan hasil sintesis dari dua lingkungan perangkat keras. Metrik performa *runtime* (RAM, TPS) merupakan pengukuran *native* pada arsitektur ARM (Tecno Pova 5), sedangkan metrik kognitif (Akurasi) diekstraksi dari arsitektur CUDA (PC RTX 3060). Sintesis ini

menggunakan asumsi ekuivalensi statistik, di mana penggunaan format *file* biner yang sama (.gguf) meminimalisir deviasi fungsional secara makroskopis.

4.4.6. Uji Statistik Signifikansi Perbedaan Performa Antar Varian (LFM 2.5)

Klaim performa pada Tabel IV.2 dilaporkan dalam bentuk rerata tiga ulangan; namun rerata saja belum cukup untuk membuktikan bahwa selisih antar varian benar-benar nyata, bukan sekadar variasi pengukuran (*noise*) yang muncul karena kondisi lingkungan eksekusi tidak pernah identik (suhu CPU, beban *background* OS, jadwal *scheduler* Android). Untuk menjawab kekhawatiran tersebut digunakan dua uji statistik. *Welch's t-test* membandingkan dua varian sekaligus dan menghitung probabilitas selisih reratanya hanya kebetulan; bila $p\text{-value} < 0,05$, selisih dinyatakan berbeda nyata (“Beda”), sementara bila $p\text{-value} \geq 0,05$ selisih dinyatakan belum cukup meyakinkan (“Tidak”). *One-way ANOVA* melengkapi pengujian dengan memeriksa apakah keempat varian secara bersamaan memang berbeda, sebelum perbandingan pasangan dilakukan. Dengan kerangka ini, klaim seperti “Q4_K_M memang lebih cepat daripada FP16” dapat dipertahankan secara inferensial, bukan hanya deskriptif.

Tabel IV. 10
Ringkasan Uji Beda Performa Antar Varian LFM 2.5

Pasangan Varian	Gen TPS	Prompt TPS	Peak RAM	Ringkasan
FP16 vs Q5_K_M	Beda	Tidak	Beda	Gen & RAM berbeda; Prompt tidak
FP16 vs Q4_K_M	Beda	Tidak	Beda	Gen & RAM berbeda; Prompt tidak
FP16 vs Q3_K_M	Beda	Tidak	Beda	Gen & RAM berbeda; Prompt tidak
Q5_K_M vs Q4_K_M	Beda	Beda	Beda	Semua metrik berbeda
Q5_K_M vs Q3_K_M	Tidak	Beda	Beda	Prompt & RAM berbeda; Gen tidak

Pasangan Varian	Gen TPS	Prompt TPS	Peak RAM	Ringkasan
Q4_K_M vs Q3_K_M	Beda	Beda	Beda	Semua metrik berbeda

Sumber: (Penulis)

Tabel IV.10 menyampaikan tiga temuan utama untuk LFM 2.5. Pertama, perbedaan *Generation Speed* dan *Peak RAM* antar varian terbukti nyata di hampir seluruh pasangan, sehingga klaim utama penelitian bahwa kuantisasi PTQ mempercepat inferensi sekaligus mereduksi konsumsi RAM tidak sekadar pola rerata yang kebetulan terjadi. Kedua, keunggulan Q4_K_M atas Q3_K_M terbukti nyata baik pada Gen TPS (13,67 vs 11,07 t/s) maupun Prompt TPS (43,67 vs 18,27 t/s); temuan ini memperkuat rekomendasi varian operasional Q4_K_M pada Sub-bab 4.4.5 dari sudut pandang murni kecepatan, bahkan sebelum mempertimbangkan retensi kepintaran. Ketiga, pengaruh kuantisasi pada *Prompt Speed* tidak homogen: pasangan FP16 vs Q3_K_M dan FP16 vs Q5_K_M tidak menunjukkan perbedaan yang meyakinkan, dan pola tidak homogen ini mengonfirmasi anomali *unpacking* bit ganjil pada Q3_K_M yang sudah dianalisis pada Sub-bab 4.4.2.

Tabel IV. 11
Ringkasan Uji Beda Performa Antar Varian Qwen 3.5

Pasangan Varian	Gen TPS	Prompt TPS	Peak RAM	Ringkasan
FP16 vs Q5_K_M	Beda	Tidak	Beda	Gen & RAM berbeda; Prompt tidak
FP16 vs Q4_K_M	Beda	Tidak	Beda	Gen & RAM berbeda; Prompt tidak
FP16 vs Q3_K_M	Beda	Beda	Beda	Semua metrik berbeda
Q5_K_M vs Q4_K_M	Tidak	Tidak	Beda	Hanya RAM berbeda
Q5_K_M vs Q3_K_M	Tidak	Beda	Beda	Prompt & RAM berbeda; Gen tidak

Pasangan Varian	Gen TPS	Prompt TPS	Peak RAM	Ringkasan
Q4_K_M vs Q3_K_M	Tidak	Beda	Beda	Prompt & RAM berbeda; Gen tidak

Sumber: (Penulis)

Keterangan: **Beda** = berbeda nyata secara statistik; **Tidak** = tidak berbeda nyata.

Tabel IV.11 mempertegas pola yang teramati pada LFM dan memperluasnya ke arsitektur *reasoning* Qwen 3.5. Pertama, reduksi *peak* RAM Qwen jauh lebih dramatis daripada LFM dan terbukti nyata pada seluruh pasangan, dengan rerata turun dari 3.744 MB (FP16) menjadi 2.856 MB (Q5_K_M), 2.563 MB (Q4_K_M), dan 1.898 MB (Q3_K_M). Kedua, peningkatan *Generation Speed* dari FP16 ke seluruh varian terkuantisasi terbukti nyata (rasio percepatan 2,28–2,65 kali lipat dari 1,87 t/s menjadi 4,25–4,95 t/s), tetapi perbedaan *Generation Speed* antar tiga varian terkuantisasi tidak nyata; pola ini menandakan Qwen 3.5 mengalami saturasi *throughput* pada CPU Helio G99 di kisaran 4–5 t/s, berbeda dengan LFM 2.5 yang masih memiliki *headroom* hingga 13,67 t/s pada Q4_K_M. Ketiga, anomali *Prompt Speed* pada Q3_K_M kembali muncul (FP16 vs Q3_K_M dan Q4_K_M vs Q3_K_M sama-sama berbeda nyata), menguatkan analisis Sub-bab 4.4.2 bahwa *bottleneck unpacking* bit ganjil pada CPU Arm bersifat arsitektur-independen dan terjadi pada kedua keluarga model selama format presisi 3-bit dipakai.

Dengan demikian, hasil uji statistik dari kedua tabel di atas memberikan dukungan kuantitatif yang konsisten dengan kerangka rekomendasi Tabel IV.8 (LFM) dan Tabel IV.9 (Qwen). Untuk LFM 2.5, Q4_K_M unggul signifikan dibanding Q3_K_M pada metrik *throughput* utama. Untuk Qwen 3.5, seluruh varian terkuantisasi memiliki RAM dan *Generation Speed* yang signifikan lebih baik dibanding FP16 tanpa perbedaan *Generation Speed* yang signifikan di antara mereka. Gabungan kedua

temuan ini menjadikan Q4_K_M sebagai varian dengan profil efisiensi paling kokoh untuk Tecno Pova 5 pada kedua arsitektur uji.

