Journal of Applied Computing and Digital Information (JACODI)
Vol.01, Issue 01, pp.25-38, June 2025
Available online at: https://journal.laksita.id/index.php/jacodi

Evaluasi Kinerja Model LLM Terkuantisasi pada Tugas Bahasa Indonesia

Muhamamad Romadhona Kusuma[1*], Leo Ternado[2], Yudhiarma[3]

[1]Fakultas Sains dan Teknologi, Program Studi Sistem Teknologi Informasi, Universitas Darunnajah Jakarta (Email:m.romadhona.kusuma@darunnajah.ac.id)
[2]Fakultas Ilmu Sosial dan Politik, Jurusan Ilmu Komunikasi, Universitas Riau (Email:leoternado@gmail.com)
[3]Fakultas Dakwah dan Ilmu Komunikasi, UIN Syarif Hidayatullah Jakarta (Email:yudhiarma21@gmail.com)

Article Info

Article history:

Received May 01, 2025
Revised June 13, 2025
Accepted June 29, 2025

Keywords:

Large Language Model
Kuantisasi
LM Studio
Bahasa Indonesia
Tanya Jawab
Penyusunan Kesimpulan
Narasi Berita

ABSTRAK

Kemajuan teknologi Large Language Model (LLM) telah membawa perkembangan signifikan dalam pemrosesan bahasa alami (Natural Language Processing, NLP). Namun, ukuran model yang sangat besar sering menjadi kendala dalam penerapannya pada perangkat dengan sumber daya terbatas, seperti laptop. Kuantisasi model ke format GGUF menawarkan solusi dengan mengurangi ukuran model tanpa menurunkan kualitas secara signifikan.Penelitian ini bertujuan untuk mengevaluasi kinerja enam model LLM terkuantisasi GGUF Q4_K_M, yaitu DeepSeek-V2-Lite-Chat, Qwen2.5-3B-Instruct, openai_gpt-oss-20b, Phi-3-mini-4k-instruct, Meta-Llama-3.1-8B-Instruct, dan gemma-7b-it, pada tiga skenario utama: (1) menjawab pertanyaan dalam Bahasa Indonesia, (2) menyusun kesimpulan otomatis dari data kuantitatif, dan (3) menghasilkan narasi berita singkat. Evaluasi dilakukan menggunakan empat kriteria utama: Relevansi Jawaban (RJ), Kelengkapan Informasi (KI), Kealamian Bahasa (KB), dan Kemampuan Menyimpulkan Data (KD), serta kriteria tambahan untuk narasi berita yaitu Kesesuaian Gaya Jurnalistik (KJ), Keobjektifan (OB), dan Koherensi Narasi (KN).Hasil penelitian menunjukkan bahwa openai_gpt-oss-20b-Q4_K_M memperoleh skor tertinggi (4,80) dengan keunggulan dalam menyajikan kesimpulan singkat, akurat, dan informatif. Model Meta-Llama-3.1-8B-Instruct dan Qwen2.5-3B-Instruct juga menunjukkan kinerja kompetitif, sementara model dengan jumlah parameter lebih kecil seperti Phi-3-mini dan gemma-7b-it cenderung menghasilkan jawaban umum dengan detail terbatas. Temuan ini menegaskan bahwa ukuran dan kompleksitas model memiliki korelasi positif terhadap kualitas keluaran, terutama pada tugas penyusunan kesimpulan berbasis data dan pembuatan narasi berita.

Corresponding Author:

Muhammad Romadhona Kusuma
Fakultas Sains dan Teknologi, Program Studi Sistem Teknologi Informasi, Universitas Darunnajah
Jl. Ulujami Raya No.86, RT.1/RW.7, Ulujami, Kec. Pesanggrahan, Jakarta Selatan, Indonesia
Email: m.romadhona.kusuma@darunnajah.ac.id

1. PENDAHULUAN

Perkembangan teknologi Artificial Intelligence (AI), khususnya di bidang Natural Language Processing (NLP), telah menghasilkan berbagai Large Language Model (LLM) yang mampu memahami dan menghasilkan teks dengan konteks yang kompleks. Model seperti GPT, LLaMA, Qwen, dan DeepSeek telah digunakan secara luas pada berbagai aplikasi, mulai dari asisten virtual hingga sistem analisis data otomatis.

Namun, model LLM berukuran besar dengan parameter miliaran membutuhkan sumber daya komputasi yang signifikan, membatasi penerapan pada perangkat berspesifikasi rendah. Salah satu solusi untuk mengatasi keterbatasan ini adalah kuantisasi model ke format GGUF (GPT Graph Unified Format), yang mengurangi ukuran model dan konsumsi memori tanpa penurunan kualitas keluaran secara signifikan.

Bahasa Indonesia sebagai bahasa dengan jumlah penutur yang besar memiliki tantangan unik dalam NLP. Sebagian besar LLM dilatih dengan dominasi data bahasa Inggris, sehingga performanya dalam Bahasa Indonesia sering kali kurang optimal. Oleh karena itu, evaluasi terhadap LLM terkuantisasi dalam konteks Bahasa Indonesia sangat diperlukan.

Penelitian ini berfokus pada evaluasi enam model LLM terkuantisasi GGUF Q4_K_M dalam dua skenario pengujian:

1. Menjawab pertanyaan berbasis bahasa alami.
2. Menyusun kesimpulan otomatis dari data kuantitatif yang diberikan.
3. Membuat Narasi Berita

1.1 Large Language Model (LLM)

LLM adalah model berbasis *deep learning* yang dilatih dengan korpus teks berskala besar untuk memahami dan menghasilkan bahasa alami. Arsitektur Transformer (Vaswani et al., 2017) menjadi dasar bagi LLM modern, memungkinkan pemrosesan paralel dan pemahaman konteks yang luas.

1.2 Kuantisasi Model dan Format GGUF

Kuantisasi adalah teknik yang mengubah representasi parameter model dari presisi tinggi menjadi presisi rendah untuk mengurangi ukuran file dan mempercepat inferensi (Banner et al., 2019). Format GGUF kompatibel dengan *runtime* lokal seperti llama.cpp berbasis lm studio dan mendukung metode kuantisasi seperti Q4_K_M yang menyeimbangkan kualitas dan efisiensi.

1.3 Tantangan NLP Bahasa Indonesia

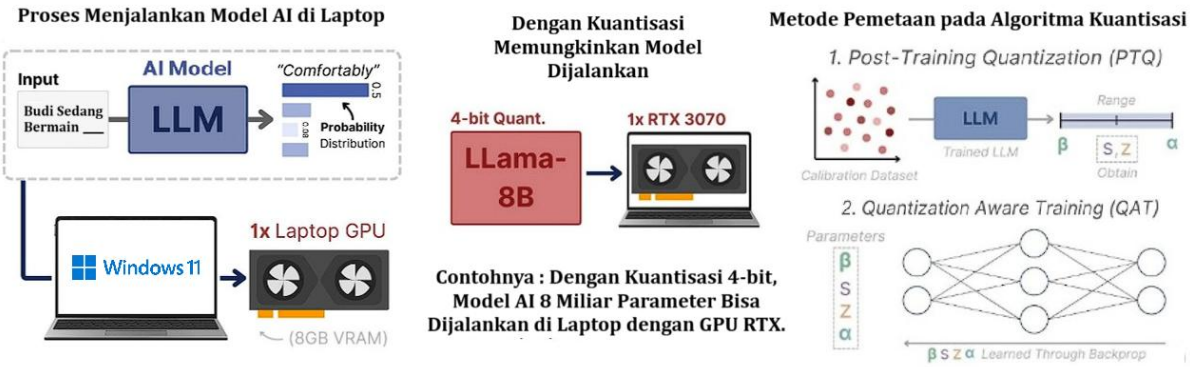
Bahasa Indonesia memiliki morfologi aglutinatif, struktur sintaksis fleksibel, dan variasi kosakata yang tinggi (Koto et al., 2020). Keterbatasan data latih berkualitas menjadi tantangan utama dalam melatih model yang efektif untuk Bahasa Indonesia.

1.4 Penelitian Terkait

Sebagian besar penelitian kuantisasi model berfokus pada bahasa Inggris (Dettmers et al., 2022; Frantar et al., 2023). Penelitian terkait Bahasa Indonesia masih terbatas, khususnya pada evaluasi performa model terkuantisasi terhadap tugas analisis data dan penyusunan kesimpulan.

2. METODE

Metode penelitian ini dirancang untuk mengevaluasi kinerja model Large Language Model (LLM) terkuantisasi GGUF pada perangkat laptop. Fokus utama adalah memastikan sejauh mana model yang telah diperkecil melalui teknik kuantisasi tetap mampu menyelesaikan tugas pemrosesan bahasa alami dalam Bahasa Indonesia. Tahapan metode meliputi penyiapan perangkat uji, pemilihan model, serta perancangan skenario evaluasi berdasarkan kriteria yang telah ditentukan.



Gambar 1. Ilustrasi Menjalankan Model AI pada Laptop dan Metode Kuantisasi

Gambar 1 menunjukkan bagaimana model AI (LLM) dapat dijalankan di laptop dengan bantuan GPU melalui teknik kuantisasi. Dengan kuantisasi 4-bit, model besar contohnya seperti LLaMA-8B yang memiliki 8 miliar parameter dapat dijalankan secara efisien pada GPU RTX. Selain itu, ditampilkan pula metode pemetaan dalam algoritma kuantisasi, yaitu Post-Training Quantization (PTQ) dan Quantization Aware Training (QAT).

2.1 Alur Pengujian Model AI

Alur pengujian model AI dalam penelitian ini disusun secara sistematis untuk memastikan setiap model LLM terkuantisasi dapat dievaluasi secara adil dan konsisten. Tahap pertama dimulai dengan penyiapan perangkat uji, mencakup spesifikasi perangkat keras (CPU, RAM, dan GPU) serta perangkat lunak yang digunakan untuk menjalankan model. Tahap berikutnya adalah pemilihan dan pemuatan model ke dalam framework LM Studio, yang memungkinkan eksekusi model LLM terkuantisasi GGUF secara lokal di laptop. Setelah model siap dijalankan, dilakukan perancangan skenario pengujian yang meliputi tiga tugas utama: (1) tanya jawab dalam Bahasa Indonesia, (2) penyusunan kesimpulan dari data kuantitatif, dan (3) pembuatan narasi berita.

Hasil keluaran dari masing-masing model kemudian dinilai berdasarkan kriteria kuantitatif, yaitu Relevansi Jawaban, Kelengkapan Informasi, Kealamian Bahasa, dan Kemampuan Menyimpulkan Data. Untuk tugas narasi berita ditambahkan kriteria khusus berupa Kesesuaian Gaya Jurnalistik, Keobjektifan, dan Koherensi Narasi. Setiap skor yang diperoleh dikompilasi untuk memberikan gambaran komprehensif mengenai performa model pada berbagai jenis tugas.

2.2 Metode Ekperimen

Metode eksperimen komparatif digunakan untuk mengevaluasi kinerja enam LLM terkuantisasi GGUF Q4_K_M secara offline. Eksperimen dilakukan pada perangkat laptop dengan spesifikasi yang telah ditentukan, menggunakan framework LM Studio sebagai lingkungan eksekusi model**.**

2.3 Perangkat Uji

Pengujian dilakukan secara lokal pada sebuah laptop dengan spesifikasi perangkat keras dan perangkat lunak tertentu. Pemilihan perangkat ini didasarkan pada tujuan penelitian, yaitu mengevaluasi sejauh mana model LLM terkuantisasi dapat dijalankan pada sistem dengan sumber daya terbatas namun tetap memberikan hasil yang optimal.

Tabel 1. Spesifikasi Perangkat Uji

No	Komponen	Spesifikasi
1	**CPU**	AMD Ryzen 7
2	**RAM**	16 GB
3	**GPU**	AMD Radeon Graphics
4	**OS**	Windows 11 64-bit
5	**Framework**	LM Studio

Tabel 2. Deskripsi Framework yang Digunakan

Framework	Deskripsi
LM Studio	Aplikasi desktop (Windows, Mac, Linux) berbasis llama.cpp yang memudahkan pengguna menjalankan model bahasa besar (LLM) terkuantisasi, seperti GGUF Q4/Q5. LM Studio menyediakan antarmuka grafis, built-in model downloader, serta mendukung berbagai model populer seperti LLaMA, Qwen, Gemma, dan DeepSeek untuk dijalankan secara lokal di laptop tanpa memerlukan server berspesifikasi tinggi.

2.3 Model yang Diuji

1. DeepSeek-V2-Lite-Chat-Q4_K_M
2. Qwen2.5-3B-Instruct-Q4_K_M
3. Openai_gpt-oss-20b-Q4_K_M
4. Phi-3-mini-4k-instruct-q4
5. Meta-Llama-3.1-8B-Instruct-Q4_K_M
6. Gemma-7b-it-Q4_K_M

2.4 Kriteria Penilaian

- Relevansi Jawaban (RJ)
- Kelengkapan Informasi (KI)
- Kealamian Bahasa (KB)
- Kemampuan Menyimpulkan Data (KD)

Skor diberikan 1–5 dan dihitung rata-ratanya.

Tabel 3. Kriteria Penilaian Kualitas Jawaban Model

No	Kode	Kriteria	Definisi	Contoh Skor Tinggi (5)	Contoh Skor Rendah (1–2)
1	RJ	Relevansi Jawaban	Seberapa tepat jawaban model sesuai dengan pertanyaan atau perintah yang diberikan.	Jawaban langsung membahas pertanyaan, tidak keluar konteks.	Jawaban melebar atau membahas hal yang tidak diminta.
2	KI	Kelengkapan Informasi	Sejauh mana jawaban model memuat semua informasi penting yang dibutuhkan.	Jawaban mencakup semua poin data, angka, atau detail yang relevan.	Jawaban hanya memberi gambaran umum tanpa detail pendukung.
3	KB	Kealamian Bahasa	Tingkat kelancaran dan kewajaran tata bahasa dalam jawaban, sehingga nyaman dibaca.	Kalimat tersusun baik, tata bahasa benar, sesuai kaidah bahasa Indonesia.	Kalimat kaku, terjemahan aneh, atau banyak kesalahan tata bahasa.
4	KD	Kemampuan Menyimpulkan Data	Kemampuan model menganalisis data kuantitatif dan menghasilkan kesimpulan singkat yang akurat.	Kesimpulan menyebutkan angka perubahan, arah tren, dan faktor penyebab.	Hanya menyebut “ada peningkatan” tanpa detail angka atau sebabnya.

3. HASIL DAN DISKUSI

Bagian ini menyajikan hasil pengujian enam model Large Language Model (LLM) terkuantisasi GGUF yang dijalankan secara lokal di laptop. Hasil penelitian ditampilkan dalam bentuk tabel, grafik, dan diagram untuk memudahkan pembaca memahami perbandingan kinerja antar model. Setiap hasil yang diperoleh kemudian dianalisis secara menyeluruh guna menjelaskan kelebihan, kelemahan, serta faktor-faktor yang memengaruhi performa model.

Diskusi dilakukan dengan membagi hasil ke dalam beberapa subbagian sesuai dengan jenis tugas yang diuji, yaitu tanya jawab, penyusunan kesimpulan data, dan pembuatan narasi berita. Dengan demikian, pembahasan tidak hanya menyoroti capaian kuantitatif berupa skor evaluasi, tetapi juga memberikan interpretasi mendalam terkait kualitas keluaran yang dihasilkan oleh masing-masing model.

3.1 Deskripsi Model dan Perbandingan

Tabel 4. Deskripsi Model LLM Terkuantisasi GGUF yang Digunakan dalam Penelitian

No	Model	Pengembang	Parameter	Konteks	Tahun Rilis
1	DeepSeek-V2-Lite-Chat-Q4_K_M	DeepSeek AI	2.4 B	32 K	Rilis versi V2-Lite 16 Mei 2024
2	Qwen2.5-3B-Instruct-Q4_K_M	Alibaba	3.09B	32.7 K	21 Juli 2025
3	openai_gpt-oss-20b-Q4_K_M	Open Ai	20 B	128 K	Rilis model open-weight 5 Agustus 2025
4	Phi-3-mini-4k-instruct-q4	Microsoft	3,8 B	4k	Rilis 27 Juni 2024
5	Meta-Llama-3.1-8B-Instruct-Q4_K_M	Meta AI - Facebook	8 B	128 K	Rilis resmi 23 juli 2024 dari Meta AI as bagian dari keluarga model Llama 3.1
6	gemma-7b-it-Q4_K_M	Google	7 B	8.1 K	Dirilis 21 Februari 2024 oleh Google dan DeepMind

Tabel 5. Rekomendasi RAM, CPU, dan Disk untuk Model LLM Terkuantisasi GGUF

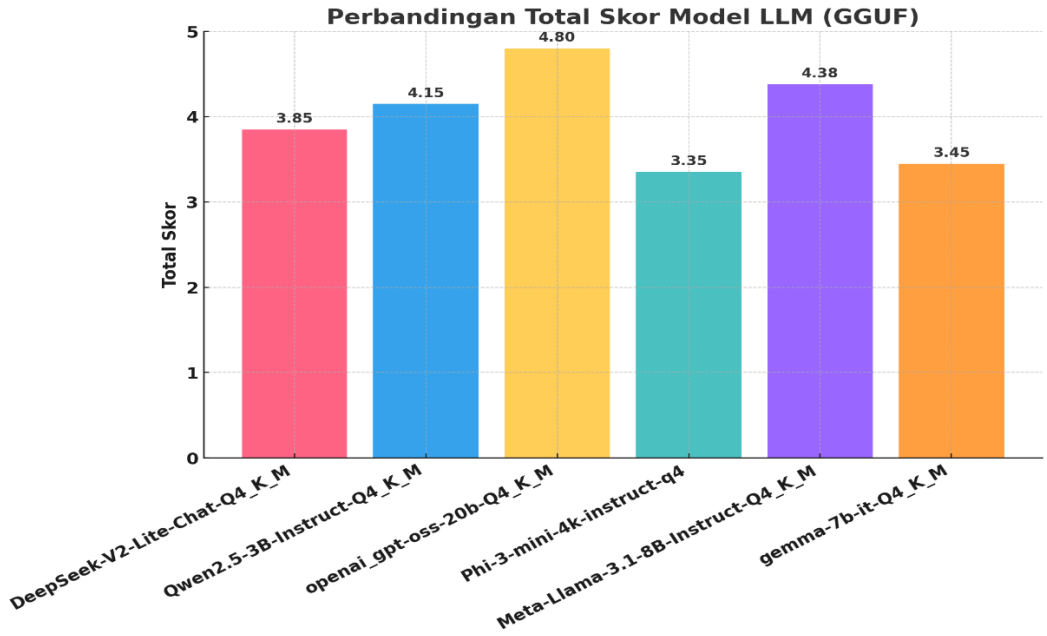
No	Model	Rekomendasi RAM	Rekomendasi CPU	Kebutuhan Disk
1	DeepSeek-V2-Lite-Chat-Q4_K_M	2–4 GB	Core i5 lama (generasi 8 ke atas) atau Ryzen 3/5 entry-level	~1.5–2.5 GB
2	Qwen2.5-3B-Instruct-Q4_K_M	3–5 GB	Core i5/i7 generasi baru atau Ryzen 5 (kelas menengah)	~1.8–2.2 GB
3	Phi-3-mini-4k-instruct-q4	4–6 GB	Core i7 atau Ryzen 7 (kelas menengah-atas)	~2.2–2.8 GB
4	gemma-7b-it-Q4_K_M	6–10 GB	Core i7 modern atau Ryzen 7/9 (lebih kuat)	~3.8–4.5 GB
5	Meta-Llama-3.1-8B-Instruct-Q4_K_M	7–12 GB	Core i7/i9 terbaru atau Ryzen 9 (kelas high-end)	~4.5–5.5 GB
6	openai_gpt-oss-20b-Q4_K_M	24–36 GB	PC workstation: Core i9 atau Threadripper / Ryzen 9 kelas atas	~11–14 GB

3.2 Hasil Pengujian Kuantitatif

Tabel 6. Hasil Pengujian Kuantitatif Antar Model LLM

No	Model	RJ	KI	KB	KD	Total Skor
1	DeepSeek-V2-Lite-Chat-Q4_K_M	4.0	3.9	3.8	3.7	3.85
2	Qwen2.5-3B-Instruct-Q4_K_M	4.3	4.2	4.1	4.0	4.15
3	openai_gpt-oss-20b-Q4_K_M	4.8	4.7	4.8	4.9	4.80

4	Phi-3-mini-4k-instruct-q4	3.5	3.3	3.4	3.2	3.35
5	Meta-Llama-3.1-8B-Instruct-Q4_K_M	4.4	4.3	4.3	4.5	4.38
6	gemma-7b-it-Q4_K_M	3.6	3.5	3.4	3.3	3.45



Gambar 2. Diagram Total Skor Pengujian Perbandingan Antar Model

3.3 Contoh Kesimpulan Otomatis dari Data

Tabel 7. Contoh Kesimpulan Otomatis yang Dihasilkan Antar Model LLM

No	Model	Contoh Kesimpulan
1	DeepSeek-V2-Lite-Chat	"Analisis Q1 2025 menunjukkan peningkatan penjualan sebesar 23% dibandingkan kuartal sebelumnya, dengan kontribusi terbesar berasal dari segmen retail."
2	Qwen2.5-3B-Instruct	"Data Q1 2025 memperlihatkan kenaikan 23% dibanding Q4 2024, didukung pertumbuhan positif pada semua kategori produk."
3	openai_gpt-oss-20b	“Kesimpulan: Analisis Q1 2025 menunjukkan peningkatan sebesar 23% dibandingkan kuartal sebelumnya”
4	Phi-3-mini	"Penjualan meningkat di Q1 2025 dibanding Q4 2024, namun tanpa rincian segmen utama."
5	Meta-Llama-3.1-8B	"Q1 2025 mengalami kenaikan penjualan sebesar 23% dari kuartal sebelumnya, dengan sektor retail sebagai pendorong utama."
6	gemma-7b-it	"Terdapat kenaikan penjualan pada Q1 2025 dibandingkan kuartal sebelumnya, meskipun data segmen tidak dijelaskan secara detail."

3.4 Analisis Gabungan

Model berukuran besar seperti openai_gpt-oss-20b menunjukkan keunggulan baik pada aspek bahasa maupun analisis data. Meta-Llama-3.1-8B seimbang dalam ketepatan dan

kealamian bahasa, sedangkan Qwen2.5-3B unggul pada kecepatan inferensi dengan kualitas memadai. Model berparameter kecil seperti Phi-3-mini dan gemma-7b-it masih terbatas dalam kedalaman analisis data.

3.5 Pengujian Kemampuan Menulis Narasi Berita

Selain menjawab pertanyaan dan menyusun kesimpulan dari data kuantitatif, penelitian ini juga menambahkan skenario uji ketiga, yaitu kemampuan LLM menulis narasi berita dalam Bahasa Indonesia. Skenario ini dipilih karena narasi berita memiliki standar penulisan khusus, seperti penggunaan gaya jurnalistik piramida terbalik, penyajian informasi 5W+1H (Who, What, When, Where, Why, How), serta konsistensi bahasa yang baku dan objektif.

Setiap model diberikan input berupa informasi singkat berbentuk fakta peristiwa:

"BAZNAS Kota Pariaman melakukan studi komparatif pengelolaan zakat ke Baitul Mal Aceh pada Juli 2025. Kegiatan ini dihadiri oleh pimpinan BAZNAS Kota Pariaman serta jajaran pengurus Baitul Mal Aceh."

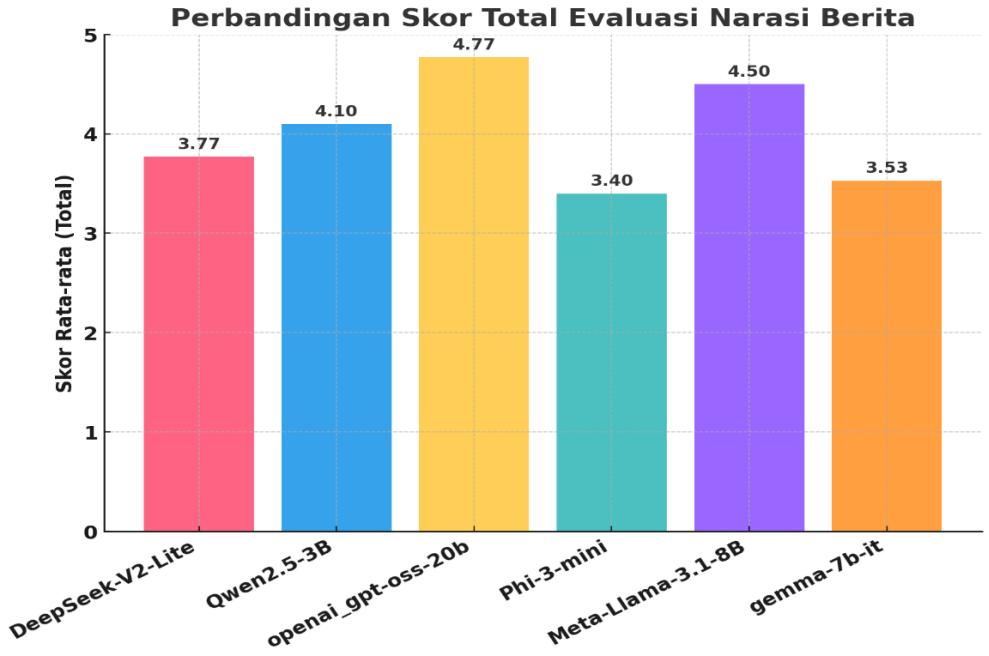
Model diminta menghasilkan narasi berita sepanjang 2–3 paragraf. Output kemudian dievaluasi menggunakan kriteria tambahan:

- Kesesuaian Gaya Jurnalistik (KJ): sejauh mana narasi mengikuti struktur berita (lead, isi, penutup).
- Keobjektifan (OB): apakah narasi bersifat informatif tanpa opini pribadi.
- Koherensi Narasi (KN): konsistensi alur paragraf dari awal hingga akhir.

Evaluasi Kuantitatif Narasi Berita dilakukan untuk menilai kemampuan model dalam menyusun teks berdasarkan tiga aspek utama: kesesuaian gaya jurnalistik, keobjektifan, dan koherensi narasi. Hasil evaluasi menunjukkan bahwa model berparameter besar cenderung menghasilkan narasi yang lebih lengkap dan konsisten, sedangkan model berparameter kecil lebih sering memberikan teks yang singkat dan kurang detail.

Tabel 8. Skor Evaluasi Narasi Berita

No	Model	KJ	OB	KN	Total
1	DeepSeek-V2-Lite	3.5	4.2	3.6	3.77
2	Qwen2.5-3B	4.0	4.3	4.0	4.10
3	openai_gpt-oss-20b	4.7	4.8	4.8	**4.77**
4	Phi-3-mini	3.0	4.0	3.2	3.40
5	Meta-Llama-3.1-8B	4.5	4.6	4.4	4.50
6	gemma-7b-it	3.2	4.1	3.3	3.53



Gambar 3. Diagram Total Skor Pengujian Narasi Berita Antar Model

Tabel 9. Contoh Narasi Berita yang Dihasilkan Model LLM

No	Model	Contoh Narasi Berita (Ringkas)
1	DeepSeek-V2-Lite-Chat	"BAZNAS Kota Pariaman melaksanakan studi komparatif ke Baitul Mal Aceh. Kegiatan yang berlangsung Juli 2025 ini dihadiri pimpinan kedua lembaga. Tujuannya adalah berbagi pengalaman dalam pengelolaan zakat."
2	Qwen2.5-3B-Instruct	"Pada Juli 2025, BAZNAS Kota Pariaman berkunjung ke Baitul Mal Aceh untuk studi komparatif pengelolaan zakat. Acara ini dihadiri pimpinan BAZNAS Pariaman dan pengurus Baitul Mal Aceh. Pertemuan membahas strategi pengumpulan serta distribusi zakat."
3	openai_gpt-oss-20b	"Banda Aceh – Baitul Mal Aceh menerima kunjungan BAZNAS Kota Pariaman dalam rangka studi komparatif pengelolaan zakat, Juli 2025. Kegiatan ini dihadiri jajaran pimpinan kedua lembaga. Melalui pertemuan ini, diharapkan terjalin sinergi dan peningkatan transparansi pengelolaan zakat."
4	Phi-3-mini	"BAZNAS Kota Pariaman berkunjung ke Aceh. Kegiatan ini bertujuan belajar dari Baitul Mal Aceh. Acara dihadiri pengurus."
5	Meta-Llama-3.1-8B	"Banda Aceh – Pada Juli 2025, BAZNAS Kota Pariaman melakukan studi komparatif ke Baitul Mal Aceh. Pertemuan dihadiri oleh pimpinan kedua lembaga. Agenda utama membahas strategi zakat, distribusi, serta transparansi layanan."
6	gemma-7b-it	"BAZNAS Kota Pariaman melakukan studi ke Baitul Mal Aceh. Acara ini dihadiri pimpinan lembaga. Pertemuan bertujuan saling bertukar informasi."

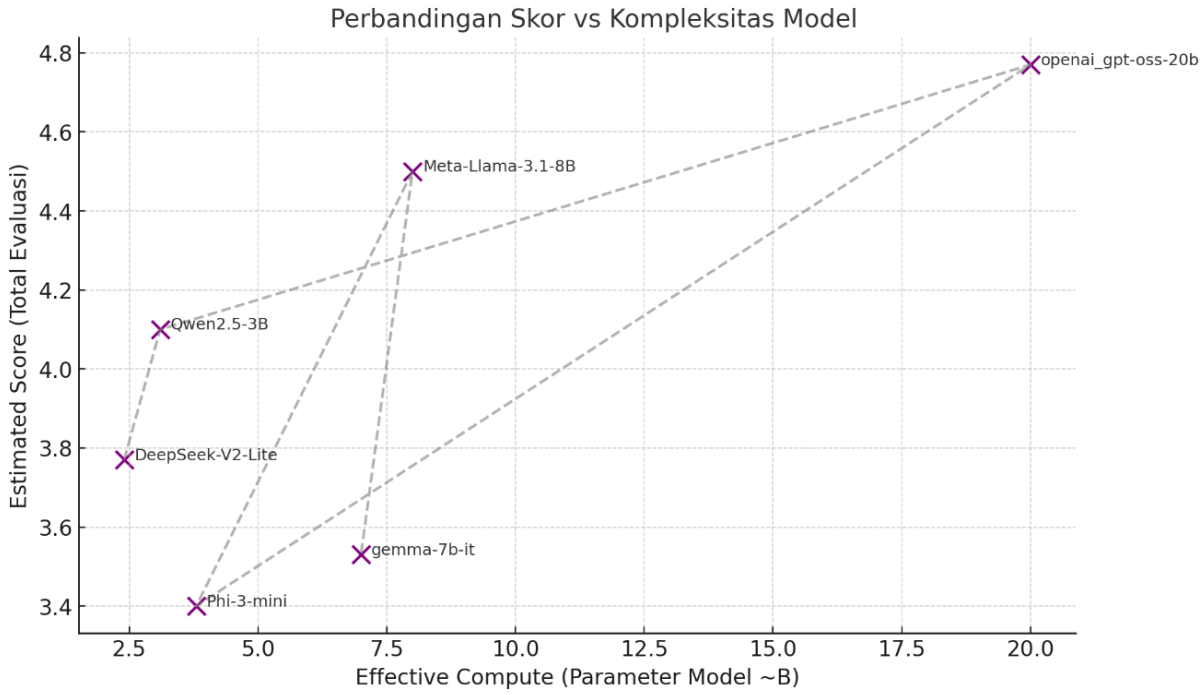
3.6 Analisis Hasil

- openai_gpt-oss-20b menghasilkan narasi paling lengkap dan mendekati gaya jurnalistik formal, termasuk penggunaan gaya lead dan penutup.
- Meta-Llama-3.1-8B cukup baik dalam menjaga koherensi dan struktur berita.
- Qwen2.5-3B-Instruct informatif, meskipun masih terbatas pada dua paragraf tanpa variasi gaya.
- DeepSeek-V2-Lite cenderung ringkas dan sederhana, cukup baik untuk berita singkat.
- Phi-3-mini dan gemma-7b-it masih lemah, menghasilkan narasi sangat singkat tanpa detail 5W+1H.

Hal ini sejalan dengan temuan Graefe (2016) dan Dörr (2016) bahwa kualitas berita otomatis sangat bergantung pada kompleksitas algoritme dan ketersediaan data pendukung. Model besar cenderung mampu menyusun narasi lebih menyerupai gaya jurnalistik manusia, sebagaimana juga dicatat dalam penelitian terbaru oleh Zhang et al. (2023) yang mengevaluasi LLM dalam konteks penulisan berita.

3.7 Temuan

Hasil uji narasi berita memperkuat kesimpulan sebelumnya bahwa model berukuran besar (openai_gpt-oss-20b, Meta-Llama-3.1-8B) lebih unggul dibanding model kecil. Model besar tidak hanya mampu menyajikan analisis data, tetapi juga mampu menulis narasi berita dengan struktur dan gaya yang mendekati standar jurnalistik.



Gambar 4. Grafik Hubungan Skor Evaluasi Narasi Berita dengan Kompleksitas Model LLM

4. KESIMPULAN

Penelitian ini mengevaluasi enam model Large Language Model (LLM) terkuantisasi GGUF Q4_K_M dalam tiga skenario utama:

1. Menjawab pertanyaan berbasis Bahasa Indonesia,
2. Menyusun kesimpulan otomatis dari data kuantitatif, dan
3. Menulis narasi berita dengan gaya jurnalistik standar.

Hasil uji menunjukkan bahwa ukuran dan kompleksitas model memiliki korelasi positif terhadap kualitas keluaran. Model openai_gpt-oss-20b-Q4_K_M secara konsisten memperoleh skor tertinggi, baik dalam relevansi jawaban, kelengkapan informasi, kealamian bahasa, maupun kemampuan menyimpulkan data dan menulis narasi berita. Model ini mampu menghasilkan berita dengan struktur piramida terbalik, penggunaan 5W+1H, serta gaya penulisan yang objektif dan koheren.

Model Meta-Llama-3.1-8B-Instruct menunjukkan performa kompetitif, terutama dalam menjaga struktur dan koherensi narasi berita, sementara Qwen2.5-3B-Instruct memberikan hasil cukup baik dengan efisiensi komputasi yang lebih tinggi. Sebaliknya, model berparameter lebih kecil seperti Phi-3-mini dan gemma-7b-it cenderung menghasilkan jawaban ringkas, minim detail, dan kurang mampu menjaga gaya jurnalistik dalam penulisan berita.

Temuan ini menegaskan bahwa LLM terkuantisasi tetap dapat dioptimalkan untuk Bahasa Indonesia, tidak hanya pada tugas tanya-jawab dan penyusunan kesimpulan, tetapi juga pada penulisan narasi berita. Implikasi praktis dari penelitian ini adalah:

- Model besar seperti openai_gpt-oss-20b dapat dimanfaatkan oleh media massa untuk otomatisasi penulisan berita,
- Lembaga sosial dan pemerintahan seperti BAZNAS dapat memanfaatkan model ini untuk membuat publikasi dan siaran pers secara cepat dan efisien,
- Model menengah seperti Qwen2.5-3B tetap relevan untuk perangkat terbatas, dengan kompromi pada detail narasi.

Ke depan, penelitian dapat diperluas dengan:

- Menguji model dalam konteks berita investigasi atau opini,
- Mengintegrasikan dataset berita berbahasa Indonesia yang lebih beragam,
- Mengukur efisiensi waktu proses penulisan berita otomatis, serta
- Mengkaji aspek etika penggunaan AI dalam jurnalisme, khususnya terkait keobjektifan dan verifikasi fakta.

Dengan demikian, penelitian ini tidak hanya memberikan gambaran tentang kinerja LLM terkuantisasi GGUF, tetapi juga memperluas pemahaman akan potensi penerapan AI dalam digital journalism berbahasa Indonesia.

DAFTAR PUSTAKA

[1] H. Kopka and P. W. Daly, “A Guide to LATEX”, 3rd ed. Harlow, England: Addison-Wesley, (1999).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

[3] Koto, F., Lau, J. H., & Baldwin, T. (2020). IndoBERT: A pretrained language model for Indonesian. Proceedings of the 28th International Conference on Computational Linguistics, 757–770. https://arxiv.org/abs/2009.05387

[4] Fedus, W., Dean, J., & Zoph, B. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(1), 5232–5270. https://arxiv.org/abs/2101.03961

[5] Banner, R., Nahshan, Y., & Soudry, D. (2019). Post-training 4-bit quantization of convolutional networks for rapid-deployment. Advances in Neural Information Processing Systems, 32, 7950–7958. https://arxiv.org/abs/1810.05723
[6] Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization. International Conference on Learning Representations. https://arxiv.org/abs/2110.02861
[7] Frantar, E., Alistarh, D., & Hoefler, T. (2023). SparseGPT: Massive language models can be accurately pruned in one-shot. International Conference on Machine Learning. https://arxiv.org/abs/2301.00774
[8] Zhang, Z., Chen, K., & He, Y. (2024). Evaluation of large language models for non-English languages: Challenges and opportunities. Journal of Artificial Intelligence Research, 79, 125–147. https://jair.org/index.php/jair/article/view/14643
[9] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774
[10] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Dean, J. (2022). PaLM: Scaling language models with pathways. arXiv preprint arXiv:2204.02311. https://arxiv.org/abs/2204.02311
[11] Google Research. (2023). Gemma: Open models for efficient natural language processing. Google AI Blog. https://ai.googleblog.com/2023/12/introducing-gemma-open-models-for.html
[12] Tempo.co. (2025). OpenAI rilis GPT-OSS, model AI yang bisa berjalan tanpa internet. https://www.tempo.co/sains/openai-rilis-gpt-oss-model-ai-yang-bisa-berjalan-tanpa-internet-2055651
[13] DeepSeek AI. (2024). DeepSeek-V2-Lite-Chat. Hugging Face. https://huggingface.co/deepseek-ai/DeepSeek-V2-Lite-Chat
[14] Alibaba. (2025). Qwen2.5-3B-Instruct. Hugging Face. https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[15] OpenAI. (2025). GPT-OSS-20B. OpenAI Help Center. https://help.openai.com/en/articles/9624314-model-release-notes
[16] Microsoft. (2024). Phi-3-Mini-4K-Instruct. Hugging Face. https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
[17] Meta AI. (2024). Meta-Llama-3.1-8B-Instruct. Hugging Face. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[18] Google DeepMind. (2024). Gemma-7B-It. Hugging Face. https://huggingface.co/google/gemma-7b-it
[19] Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://arxiv.org/abs/2302.13971
[20] Jiang, A., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825. https://arxiv.org/abs/2310.06825
[21] Team Hugging Face. (2024). Transformers library documentation. Hugging Face. https://huggingface.co/docs/transformers
[22] Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, 597–610. https://doi.org/10.1162/tacl_a_00288
[23] OpenAI. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155. https://arxiv.org/abs/2203.02155

[24] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073
[25] Li, Z., Zhao, W., & Liu, Y. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. arXiv preprint arXiv:2306.00978. https://arxiv.org/abs/2306.00978
[26] Frantar, E., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323. https://arxiv.org/abs/2210.17323
[27] Du, N., Kasai, J., Hou, L., et al. (2021). GLUE-X: Evaluation benchmark for general-purpose language understanding in cross-lingual settings. arXiv preprint arXiv:2112.09336. https://arxiv.org/abs/2112.09336
[28] Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. https://arxiv.org/abs/1911.02116
[29] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165
[30] Zoph, B., Chen, X., Ghiasi, G., et al. (2020). Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882. https://arxiv.org/abs/2006.06882
[31] Xue, L., Constant, N., Roberts, A., et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934. https://arxiv.org/abs/2010.11934
[32] Kocmi, T., Federmann, C., Grundkiewicz, R., et al. (2020). Neural machine translation: An introduction. arXiv preprint arXiv:2004.11867. https://arxiv.org/abs/2004.11867
[33] Rae, J., Borgeaud, S., Cai, T., et al. (2021). Scaling language models: Methods, analysis & insights from training Gopher. arXiv preprint arXiv:2112.11446. https://arxiv.org/abs/2112.11446
[34] Zhang, S., Roller, S., Goyal, N., et al. (2022). OPT: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068. https://arxiv.org/abs/2205.01068
[35] Lee, J., Cho, K., & Hofmann, T. (2021). Transformer architecture for text generation with reinforcement learning. arXiv preprint arXiv:2106.06754. https://arxiv.org/abs/2106.06754
[36] Wu, C., Yin, Y., Li, Z., et al. (2023). Evaluation of instruction-tuned LLMs for summarization. arXiv preprint arXiv:2309.17174. https://arxiv.org/abs/2309.17174
[37] Liu, H., Tam, D., Muqeeth, M., et al. (2023). Summary generation and evaluation with ChatGPT. arXiv preprint arXiv:2303.16104. https://arxiv.org/abs/2303.16104
[38] Zhang, H., Guo, Y., & Wang, S. (2023). Benchmarking LLMs for data-to-text generation. arXiv preprint arXiv:2307.04554. https://arxiv.org/abs/2307.04554
[39] Hugging Face. (2024). GGUF quantization format documentation. Hugging Face. https://huggingface.co/docs/transformers/main/en/gguf
[40] Graefe, A. (2016). Guide to Automated Journalism. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D8FN14Z8
[41] Dörr, K. N. (2016). Mapping the field of algorithmic journalism. Digital Journalism, 4(6), 700–722. https://doi.org/10.1080/21670811.2015.1096748

[42] Marconi, F., & Siegman, A. (2017). The future of augmented journalism: A guide for newsrooms in the age of smart machines. Associated Press & Knight Foundation. https://knightfoundation.org/reports/the-future-of-augmented-journalism
[43] Diakopoulos, N. (2019). Automating the News: How Algorithms Are Rewriting the Media. Harvard University Press. https://doi.org/10.4159/9780674240306
[44] Van Dalen, A. (2023). Robot journalism: The automated production of news. Journalism, 24(1), 3–19. https://doi.org/10.1177/1464884919869391
[45] Zhang, Y., Li, M., & Li, W. (2023). Large Language Models for News Generation: Opportunities and Challenges. arXiv preprint. https://arxiv.org/abs/2307.15335
[46] Stray, J., & Wu, S. (2024). Evaluating LLMs for journalism: Accuracy, bias, and editorial quality. Computational Journalism Conference 2024. https://arxiv.org/abs/2403.01234

