Technologia : Jurnal Ilmiah
p-ISSN: 2086-6917, e-ISSN: 2656-8047
Homepage: https://ojs.uniska-bjm.ac.id/index.php/JIT/index
Vol. 17, No. 4, Oktober 2026

LOCAL LARGE LANGUAGE MODEL (LLM) UNTUK ANALISIS NARATIF PADA DASHBOARD BUSINESS INTELLIGENCE BERBASIS PRIVASI

Rudi Hartono[1*], Aso Sudiarjo[2], Shinta Siti Sundari[3]
[1,2,3] Program Studi Teknik Informatika, Fakultas Teknik,UNPER, Indonesia.

Informasi Artikel:
Dikirim: 28-09-2026; Diterima: 30-09-2026; Diterbitkan: 20-10-2026
Doi : http://dx.doi.org/10.31602/tji.v17i4.25255

Abstrak

Dashboard Business Intelligence (BI) umumnya menyajikan data dalam bentuk visual tanpa penjelasan naratif mengenai kondisi bisnis. Penelitian ini bertujuan membandingkan kinerja tiga *Large Language Models* (LLM) sumber terbuka yang dijalankan secara lokal, yaitu Llama 3 (8B), Qwen 2.5 (7B), dan Mistral (7B), dalam menghasilkan narasi wawasan bisnis dari data dashboard. Penelitian menggunakan eksperimen kuantitatif dengan tiga skenario bisnis sintetis, yaitu pertumbuhan, krisis pelanggan, dan tekanan margin. Setiap skenario diuji lima kali pada server lokal dengan memori 16 GB, sehingga menghasilkan 45 eksekusi. Kualitas narasi dievaluasi berdasarkan kesesuaian *n-gram* dan ketepatan informasi numerik, yaitu recall, presisi, dan F1 entitas numerik, serta penilaian rubrik oleh dua penilai independen, sedangkan efisiensi diukur melalui *throughput*, latensi, dan penggunaan memori. Hasil menunjukkan bahwa Llama 3 memiliki throughput tertinggi sebesar 9,22 token/detik, recall numerik 80,62%, serta presisi numerik tertinggi 94,30% (F1 86,28%). Qwen 2.5 memperoleh skor rubrik tertinggi (4,60 dari 5,00, weighted kappa 0,78), sedangkan Mistral menggunakan memori paling rendah sebesar 11,92 GB. Hasil ini menunjukkan bahwa evaluasi narasi bisnis perlu mempertimbangkan ketepatan informasi dan efisiensi, tidak hanya kesamaan kata dengan teks referensi.

Keywords: *Business intelligence, Local* LMM, *Llama, Qwen, Mistral*

This is an open-access article under a Creative Commons Attribution 4.0 International (CC-BY 4.0) License. Copyright ©**2026** by authors.

Pendahuluan

Pemantauan kinerja bisnis banyak menggunakan dashboard *Business Intelligence* (BI) untuk menyajikan berbagai indikator kinerja utama (*Key Performance Indicators/KPI*) dalam bentuk visual interaktif [5]. Namun, dashboard umumnya hanya menampilkan angka dan grafik tanpa menjelaskan makna dari data tersebut. Pengguna masih perlu melakukan analisis secara mandiri untuk memahami kondisi bisnis. Narasi bisnis setidaknya perlu menjelaskan kondisi kinerja, penyebab perubahan atau masalah, serta tindakan strategis yang dapat dilakukan. Upaya mengubah data menjadi teks (*data-to-text*) telah lama dikembangkan dalam bidang *Natural Language Processing* (NLP) [9], [16], tetapi pendekatan berbasis templat cenderung kaku dan membutuhkan pemeliharaan ketika struktur data berubah.

Perkembangan arsitektur *transformer* [18] dan *Large Language Models* (LLM) [1], [22] memungkinkan model menghasilkan teks berdasarkan instruksi manusia [14], termasuk narasi

analitis dari data dashboard. Namun, penggunaan LLM berbasis *cloud* menimbulkan tantangan ketika data bisnis bersifat sensitif, seperti omzet, margin, dan informasi pelanggan. Data tersebut perlu dikirim ke server pihak ketiga sehingga berkaitan dengan perlindungan data dan kepatuhan terhadap regulasi, seperti UU Pelindungan Data Pribadi [17] dan *General Data Protection Regulation* (GDPR) [7]. Salah satu alternatifnya adalah menggunakan LLM [2] sumber terbuka secara lokal (*on-premise*), sehingga data tetap berada di lingkungan organisasi dan biaya operasional dapat lebih terkendali.

Berbagai model *open-weights* berukuran 7–8 miliar parameter kini dapat dijalankan pada perangkat keras umum setelah proses kuantisasi [8], seperti Llama 3 [6], Mistral 7B [11], dan Qwen 2.5 [20]. Meskipun memiliki ukuran yang relatif sama, ketiganya dikembangkan dengan arsitektur, data pelatihan, dan kemampuan bahasa yang berbeda. Berdasarkan penelusuran penulis, penelitian yang secara khusus membandingkan ketiga model tersebut untuk menghasilkan narasi BI berbahasa Indonesia masih terbatas. Perbandingan yang mencakup kualitas teks, ketepatan angka, dan efisiensi sumber daya secara bersamaan juga masih jarang ditemukan. Hal ini penting karena kesalahan angka dalam laporan bisnis dapat memengaruhi interpretasi dan keputusan, sementara *halusinasi* masih menjadi tantangan dalam penggunaan LLM [10].

Evaluasi kualitas narasi juga memiliki tantangan tersendiri. Metrik berbasis *n-gram*, seperti BLEU [15] dan ROUGE [12], masih banyak digunakan, tetapi hasilnya belum selalu sejalan dengan penilaian manusia, terutama pada teks yang kompleks [3], [4]. Metrik berbasis semantik [21] dan penilaian menggunakan LLM [13], [23] dapat menjadi alternatif, tetapi membutuhkan sumber daya komputasi lebih besar dan belum secara khusus memastikan ketepatan angka. Oleh karena itu, evaluasi narasi bisnis perlu mempertimbangkan ketepatan informasi numerik, terutama dalam memastikan angka penting dari dashboard disampaikan secara lengkap dan akurat.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan: (1) membandingkan kualitas teks dan ketepatan angka dari narasi yang dihasilkan Llama 3 (8B), Qwen 2.5 (7B), dan Mistral (7B), (2) membandingkan efisiensi ketiga model berdasarkan latensi, *throughput*, dan penggunaan memori pada server lokal 16 GB, serta (3) mengidentifikasi karakteristik model yang sesuai untuk penerapan LLM lokal pada dashboard BI.

Penelitian ini memberikan tiga kontribusi utama, yaitu protokol eksperimen dengan pengujian berulang dan penilaian rubrik oleh penilai independen, metrik recall, presisi, dan F1 numerik yang menyesuaikan format angka Indonesia, serta pedoman pemilihan model dan perancangan arsitektur LLM lokal untuk sistem BI. Hasil penelitian diharapkan dapat menjadi referensi bagi organisasi dalam menerapkan pelaporan bisnis berbasis narasi secara otomatis.

Tinjauan Pustaka

Business Intelligence (BI) merupakan sistem yang mengolah dan menyajikan data untuk mendukung pengambilan keputusan organisasi [5]. Dalam *Natural Language Generation* (NLG), pembentukan narasi meliputi pemilihan informasi, penyusunan struktur, dan realisasi teks [16]. Perkembangan *Large Language Models* (LLM) memungkinkan proses tersebut dilakukan secara lebih fleksibel melalui *prompt* tanpa bergantung pada aturan atau templat yang kaku [1], [14].

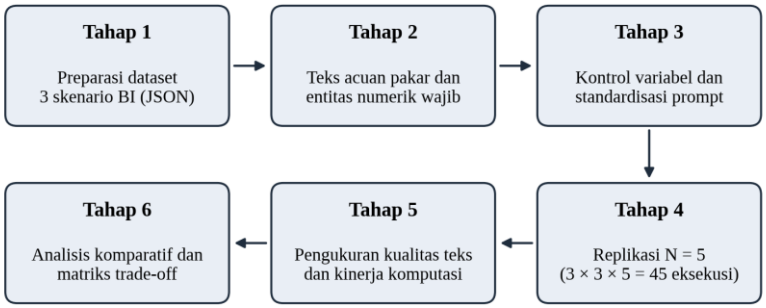
Arsitektur *transformer* dengan mekanisme *attention* menjadi dasar berbagai LLM modern [18], [6], [11], [20]. Untuk menjalankan model pada perangkat dengan sumber daya terbatas, kuantisasi dapat digunakan untuk mengurangi ukuran bobot dan kebutuhan memori dengan penurunan kualitas yang relatif kecil [8]. Format GGUF dan mesin inferensi seperti Ollama mendukung penggunaan model terkuantisasi secara lokal.

Evaluasi kualitas narasi dapat menggunakan BLEU dan ROUGE yang mengukur kesamaan *n-gram* antara teks hasil dan teks acuan [12], [15]. Namun, kedua metrik tersebut memiliki keterbatasan dalam mengenali parafrase dan memastikan kebenaran informasi [3]. Metrik berbasis semantik seperti BERTScore [21] serta pendekatan G-Eval dan *LLM-as-a-judge* [13], [23] dapat memberikan penilaian yang lebih luas, tetapi membutuhkan sumber daya lebih

besar. Pada narasi BI, ketepatan angka menjadi aspek penting yang dapat diperiksa secara langsung terhadap data asli. Oleh karena itu, recall, presisi, dan F1 entitas numerik digunakan sebagai pelengkap metrik kesamaan teks.

Metodologi

Penelitian ini menggunakan metode eksperimen kuantitatif untuk menguji kinerja model bahasa besar (LLM) secara empiris. Pengujian dilakukan pada server lokal (*on-premise*) dengan kapasitas memori 16 GB menggunakan Ollama sebagai mesin inferensi. Seluruh proses inferensi dijalankan menggunakan CPU. Spesifikasi lingkungan: prosesor AMD Ryzen 5 5600G (6 inti/12 utas, 3,9 GHz, boost hingga 4,4 GHz, cache L3 16 MB), RAM DDR4-3200 16 GB dual-channel, SSD NVMe 512 GB, Windows 11 Pro 64-bit (versi 10.0.26200), Ollama v0.34.4 (backend llama.cpp dengan AVX2), Python 3.14.4, NLTK 3.10.3, rouge-score 0.1.2, psutil 7.2.2, dan pandas 3.0.2. Model yang diuji terdiri dari tiga LLM sumber terbuka, yaitu Meta Llama 3 (8B), Alibaba Qwen 2.5 (7B), dan Mistral AI (7B). Secara keseluruhan, proses penelitian dilakukan melalui enam tahap utama seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

Dataset skenario bisnis

Untuk mengurangi ketergantungan pada satu jenis dataset, penelitian ini menggunakan tiga skenario bisnis yang dirancang agar menyerupai kondisi operasional nyata. Setiap skenario dibuat dalam format JSON dan mewakili kondisi bisnis yang berbeda (Tabel 1). Di dalam setiap skenario terdapat data metrik dashboard, teks acuan sebagai pembanding, serta daftar informasi numerik yang wajib disampaikan dalam narasi.

Tabel 1. Skenario Bisnis yang Digunakan dalam Pengujian

No	Skenario	Tujuan Uji	Metrik Kunci
1	Pertumbuhan penjualan *e-commerce*	Kemampuan mengenali keberhasilan ekspansi	Omzet Rp 1.450.000.000 (target Rp 1.300.000.000, capaian 111,5%, naik 18,5% dari bulan lalu), 12.400 pesanan (+12,0%), rata-rata nilai pesanan Rp 116.935, konversi 3,4%, biaya akuisisi pelanggan Rp 42.000 (−8,7%), kontribusi Elektronik 42,8% dan Fashion 29,7%, kampanye akhir bulan menyumbang 35% volume.
2	Krisis layanan SaaS dan *churn* pelanggan	Sensitivitas terhadap krisis dan analisis akar masalah	Pendapatan berulang bulanan Rp 320.000.000 (−7,2%), *churn* 4,8% (ambang aman 2,5%, bulan lalu 1,9%), NPS turun dari 54 ke 28, respons tiket 6,4 jam (standar 2,0 jam), 348 tiket aktif, 42 *bug* kritis pasca rilis v2.4 (58% alasan *churn*).
3	Efisiensi logistik dan tekanan margin ritel	Deteksi anomali tersembunyi	Penjualan kotor Rp 2.800.000.000 (+1,2%), beban logistik Rp 610.000.000 (+22,4%), margin kotor turun dari 28,5% ke 22,1%, pengiriman tepat waktu 84,2% (target 95,0%), biaya bahan bakar dan lembur kurir naik 34,0%, rasio retur 5,6%.

Teks acuan dan rubrik validasi

Untuk mengukur kinerja metrik *n-gram*, setiap skenario dilengkapi dengan teks referensi yang disusun berdasarkan rubrik analisis bisnis eksekutif. Rubrik tersebut mencakup tiga

komponen utama, yaitu: (1) ringkasan kondisi bisnis dan metrik utama, (2) analisis penyebab atau akar masalah, dan (3) dua hingga empat rekomendasi strategis. Setiap teks referensi diperiksa kembali untuk memastikan bahwa seluruh informasi dan angka yang digunakan sesuai dengan data asli pada dashboard. Dengan demikian, teks referensi memiliki tingkat kebenaran kuantitatif sebesar 100%.

Selain teks referensi, data dashboard juga digunakan untuk menentukan informasi numerik yang wajib muncul dalam narasi, yang menjadi dasar perhitungan *recall*, presisi, dan F1 numerik dari narasi yang dihasilkan oleh masing-masing model.

Model dan konfigurasi inferensi

Ketiga model dijalankan menggunakan Ollama dalam format GGUF yang telah melalui proses kuantisasi (Tabel 2). Agar hasil perbandingan lebih adil, seluruh kondisi pengujian di luar model dibuat sama. Nilai *temperature* ditetapkan sebesar 0,2 untuk mengurangi variasi hasil dan risiko kesalahan dalam menghasilkan angka. Nilai *seed* menggunakan 42 + *trial*, sehingga setiap pengujian tetap menghasilkan variasi, tetapi prosesnya dapat direproduksi. Parameter lain: *top_p* 0,90, *top_k* 40, *repeat_penalty* 1,10, *num_thread* 12, jendela konteks (*num_ctx*) 2.048 token (prompt masukan 176–284 token), dan batas keluaran *num_predict* = −1 (generasi berhenti pada token akhir model). Mode *stream* dinonaktifkan agar waktu pemrosesan dapat diukur dengan lebih akurat. Seluruh model juga diuji menggunakan pendekatan *zero-shot*, yaitu model menghasilkan narasi tanpa melalui proses pelatihan atau *fine-tuning* khusus untuk tugas tersebut.

Semua model diberikan *system prompt* yang sama agar instruksi pengujian tetap konsisten. Dalam prompt tersebut, model ditetapkan sebagai analis Business Intelligence senior dan diminta untuk: (i) menjelaskan tren serta pencapaian target menggunakan angka yang tepat, (ii) menjelaskan penyebab terjadinya anomali, (iii) memberikan rekomendasi strategis yang dapat diterapkan, dan (iv) menggunakan bahasa Indonesia yang formal dan ringkas. Sebelum pengujian utama dilakukan, setiap model menjalani satu kali proses *warm-up*. Proses ini bertujuan untuk mengurangi pengaruh waktu yang dibutuhkan model untuk memuat bobot dari penyimpanan ke memori (*cold-start*), sehingga pengukuran latensi pada pengujian utama dapat dilakukan secara lebih akurat.

Tabel 2. Konfigurasi Model yang Diuji

Model	Tag Ollama	Parameter	Kuantisasi	Ukuran berkas
Meta Llama 3	llama3:8b-instruct-q4_0	8,03B	Q4_0	4,66 GB
Alibaba Qwen 2.5	qwen2.5:7b-instruct-q4_K_M	7,61B	Q4_K_M	4,68 GB
Mistral AI	mistral:7b-instruct-v0.3-q4_K_M	7,25B	Q4_K_M	4,37 GB

Digest SHA-256 (12 karakter awal): Llama 3 365c0bd3c000, Qwen 2.5 845dbda0ea48, Mistral 6577803aa9a0. Alias *latest* dan 7b yang dipakai saat pengujian merujuk digest tersebut, Mistral yang digunakan adalah versi v0.3.

Metrik evaluasi

Kualitas narasi diukur dengan tiga kelompok metrik otomatis. Pertama, BLEU [15] dihitung dengan pustaka NLTK menggunakan tokenisasi kata dan fungsi penghalusan metode 1 [4] untuk BLEU-1, BLEU-2, dan BLEU-4:

$$\text{BLEU-N} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right)$$

Dengan BP adalah penalti kependekan, w_n bobot tiap orde *n*-gram, dan p_n presisi *n*-gram termodifikasi. Kedua, ROUGE [12] dihitung dengan pustaka *rouge-score* tanpa *stemmer* bahasa Inggris agar sesuai dengan morfologi bahasa Indonesia, mencakup ROUGE-1, ROUGE-2, dan ROUGE-L dalam ukuran F1. ROUGE-L didefinisikan pada Persamaan (2), dengan R_{LCS} dan P_{LCS}

masing-masing *recall* dan presisi subsekuens umum terpanjang serta β pembobot relatif keduanya.

$$\text{ROUGE-L}_{\text{F1}} = \frac{(1+\beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}$$

Ketiga, integritas data kuantitatif dinilai dengan *recall*, presisi, dan F1 entitas numerik, yaitu metrik yang dirancang dalam penelitian ini. Dari setiap skenario diekstraksi himpunan *M* entitas numerik wajib $E = \{e_1, \ldots, e_M\}$, lalu keberadaan setiap entitas diperiksa dalam narasi model (hipotesis *H*) dengan toleransi format angka Indonesia: pemisah desimal titik dan koma disetarakan (misalnya 18.5% dengan 18,5%), demikian pula pemisah ribuan (12.400 dengan 12400 atau 12,400). *Recall* dihitung sebagai rasio perolehan entitas (*factual entity recall*):

$$\text{Akurasi Faktual (\%)} = \left(\frac{\sum_{i=1}^{M} I(e_i \in H)}{M}\right) \times 100$$

dengan $I(\cdot)$ sebagai fungsi indikator. Dengan TP entitas wajib yang cocok, FN entitas wajib yang tidak muncul, dan FP angka narasi yang tidak bersumber dari dashboard, presisi = TP/(TP + FP) × 100 dan F1 = 2 × presisi × *recall*/(presisi + *recall*). Pembulatan tidak dianggap cocok (112% terhadap 111,5% menjadi FP sekaligus FN), angka negatif dicocokkan bersama kata arah tren dan pembalikan arah dihitung FP, "persen" setara dengan %, Rp/IDR serta juta/miliar diekspansi ke angka penuh, angka berulang dihitung sekali, angka mirip tetapi salah (12.450 terhadap 12.400) menjadi FN sekaligus FP, angka karangan di luar data, penomoran butir, dan tahun menjadi FP.

Efisiensi komputasi diekstraksi dari metadata mesin Ollama. *Throughput* dinyatakan dalam token per detik (*tokens per second*, TPS) menurut Persamaan (4), dengan n_{eval} jumlah token yang dibangkitkan dan t_{eval} durasi pembangkitan dalam nanodetik. Total latensi diukur dalam detik sejak permintaan HTTP dikirim hingga respons lengkap diterima. Konsumsi memori puncak (*peak* RAM) dicatat dari memori fisik sistem yang terpakai sebelum dan selama inferensi menggunakan pustaka *psutil*. Selain itu, jumlah token prompt (*prompt_eval_count*) dan keluaran (*eval_count*), waktu pemuatan model, waktu evaluasi prompt, dan waktu generasi dicatat secara terpisah dari metadata yang sama.

$$\text{TPS} = \frac{n_{\text{eval}}}{t_{\text{eval}} \times 10^{-9}}$$

Replikasi dan analisis data

Pengujian tunggal tidak memadai secara statistik karena proses latar belakang sistem operasi bersifat dinamis. Oleh sebab itu setiap kombinasi model dan skenario dijalankan lima kali (N = 5), sehingga total eksekusi adalah 3 model × 3 skenario × 5 replikasi = 45, atau 15 eksekusi per model. Lima replikasi pada prompt dan data yang sama adalah replikasi teknis, bukan observasi bisnis independen, variasi bisnis diwakili tiga skenario. Kontrol seragam untuk semua model: *seed* 42 + indeks replikasi (42–46), eksekusi berurutan model demi model per skenario, satu *warm-up* dengan prompt sama (waktu muat model tetap tercatat pada replikasi pertama, Tabel 5), serta proses latar belakang non-esensial dinonaktifkan. Setiap metrik dilaporkan sebagai rata-rata sampel (Persamaan 5) dan simpangan baku sampel (Persamaan 6) dengan format $\bar{X} \pm s$.

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}$$

Seluruh data dikumpulkan dan diproses otomatis oleh skrip *benchmark* ke dalam berkas CSV. Analisis dilakukan melalui tiga cara: (i) komparasi kuantitatif nilai $\bar{X} \pm s$ antarmodel secara agregat dan per skenario, (ii) penilaian rubrik kualitatif oleh dua penilai independen, serta (iii) analisis *trade-off* antara sumber daya dan kualitas untuk menghasilkan rekomendasi implementasi. Mengingat jumlah replikasi yang kecil, perbedaan antarmodel ditafsirkan secara deskriptif dengan membandingkan selisih rata-rata terhadap simpangan baku, tanpa uji signifikansi.

Penilaian rubrik kualitatif

Sembilan narasi (3 model × 3 skenario) dipilih acak, dianonimkan, dan diurutkan acak berbeda oleh penilai independen (spesialis data dan BI senior >6 tahun, konsultan strategi manajemen >7 tahun). Penilaian mandiri berskala Likert 1–5 pada empat dimensi: struktur, kelengkapan analisis dan diagnosis, relevansi rekomendasi, serta keterbacaan dan nada eksekutif. Kesepakatan dihitung dengan Cohen's weighted kappa berbobot kuadratik [24] dan ditafsirkan menurut Landis dan Koch [25].

Hasil

Tabel 3 merangkum rata-rata kinerja ketiga model pada 45 eksekusi, sedangkan Tabel 4 menyajikan rata-rata beserta simpangan baku untuk metrik utama. Ringkasan visual disajikan pada Gambar 2. Dekomposisi token dan waktu disajikan pada Tabel 5, hasil per skenario pada Tabel 6, dan skor rubrik pada Tabel 7.

Tabel 3. Ringkasan Rata-Rata Kinerja Model (N = 5 Replikasi)

Model	TPS ↑	Latensi (s) ↓	RAM puncak (GB) ↓	ROUGE-1 F1 ↑	ROUGE-2 F1 ↑	ROUGE-L F1 ↑	BLEU-1 ↑	BLEU-4 ↑	Recall numerik ↑
Meta Llama 3 (8B)	**9,22**	**65,08**	12,22	**0,386**	0,149	**0,243**	**0,272**	0,041	**80,62%**
Alibaba Qwen 2.5 (7B)	8,99	88,00	12,73	0,378	**0,155**	0,220	0,264	**0,052**	79,21%
Mistral AI (7B)	8,92	92,24	**11,92**	0,370	0,139	0,228	0,269	0,036	76,81%

Keterangan: cetak tebal menunjukkan nilai terbaik pada setiap kolom, ↑ semakin tinggi semakin baik, ↓ semakin rendah semakin baik.

Tabel 4. Evaluasi Statistik (Rata-Rata ± Simpangan Baku)

Model	Throughput (token/detik)	Total latensi (detik)	Recall numerik (%)	Presisi numerik (%)	F1 numerik (%)
Meta Llama 3 (8B)	9,22 ± 0,19	65,08 ± 15,82	80,62 ± 12,21	94,30 ± 8,83	86,28 ± 8,22
Alibaba Qwen 2.5 (7B)	8,99 ± 0,38	88,00 ± 14,07	79,21 ± 13,95	85,93 ± 8,47	81,97 ± 10,05
Mistral AI (7B)	8,92 ± 0,26	92,24 ± 13,74	76,81 ± 18,17	88,25 ± 11,28	80,68 ± 12,81

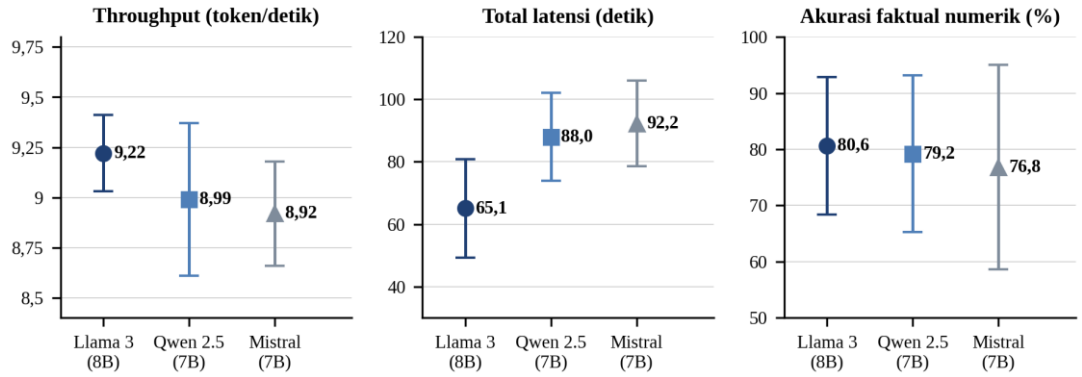
Efisiensi komputasi. *Throughput* ketiga model berada pada rentang yang cukup dekat, yaitu 8,92–9,22 token/detik, dengan simpangan baku antara 0,19–0,38. Llama 3 menghasilkan *throughput* tertinggi dengan waktu latensi total paling rendah, yaitu 65,08 detik. Sementara itu, Qwen 2.5 membutuhkan waktu 88,00 detik dan Mistral 92,24 detik. Dari sisi penggunaan memori, Mistral memiliki konsumsi memori puncak paling rendah, yaitu 11,92 GB. Selanjutnya,

penggunaan memori Llama 3 mencapai 12,22 GB, sedangkan Qwen 2.5 mencapai 12,73 GB. Dekomposisi lengkap disajikan pada Tabel 5.

Tabel 5. Dekomposisi Token dan Waktu Inferensi
(Rata-Rata ± Simpangan Baku, N = 15 per Model)

Metrik	Meta Llama 3 (8B)	Alibaba Qwen 2.5 (7B)	Mistral AI (7B)
Token prompt	234,00 ± 44,38	234,00 ± 44,38	234,00 ± 44,38
Token keluaran	521,73 ± 70,16	714,80 ± 105,79	740,93 ± 108,61
Waktu muat model, replikasi 1 / 2–5 (s)	18,17 ± 3,39 / 0,00	17,49 ± 4,69 / 0,00	22,28 ± 4,12 / 0,00
Evaluasi prompt (s)	0,106 ± 0,003	0,108 ± 0,005	0,110 ± 0,005
Waktu generasi (s)	56,74 ± 8,52	79,43 ± 10,74	83,21 ± 12,59
TPS (token/detik)	9,22 ± 0,19	8,99 ± 0,38	8,92 ± 0,26
Total latensi (s)	65,08 ± 15,82	88,00 ± 14,07	92,24 ± 13,74
Latensi ternormalisasi 500 token (s)	54,34	55,73	56,16

Kualitas teks. Berdasarkan hasil evaluasi, Llama 3 memiliki nilai rata-rata tertinggi pada ROUGE-1 sebesar 0,386, ROUGE-L sebesar 0,243, dan BLEU-1 sebesar 0,272. Sementara itu, Qwen 2.5 memperoleh nilai tertinggi pada ROUGE-2 sebesar 0,155 dan BLEU-4 sebesar 0,052. Untuk *recall* numerik, Llama 3 memperoleh nilai rata-rata sebesar 80,62 ± 12,21%, diikuti Qwen 2.5 sebesar 79,21 ± 13,95%, dan Mistral sebesar 76,81 ± 18,17%. Hasil ini menunjukkan bahwa Llama 3 memiliki rata-rata *recall* numerik tertinggi di antara ketiga model, meskipun terdapat variasi hasil pada setiap pengulangan pengujian. Llama 3 juga memiliki presisi numerik tertinggi (94,30 ± 8,83%), diikuti Mistral (88,25 ± 11,28%) dan Qwen 2.5 (85,93 ± 8,47%), sehingga F1 numerik Llama 3 (86,28%) lebih tinggi daripada Qwen 2.5 (81,97%) dan Mistral (80,68%). Simpangan baku ROUGE-L berkisar 0,04–0,06 dan BLEU-4 0,01–0,02. Hasil per skenario disajikan pada Tabel 6 (S1: pertumbuhan e-commerce, S2: krisis churn SaaS, S3: tekanan margin ritel).



Gambar 2. Perbandingan Kinerja Model (Rata-Rata ± Simpangan Baku)

Tabel 6. Kinerja per Skenario Bisnis (Rata-Rata ± Simpangan Baku, N = 5)

Skenario · Model	Token keluaran	Latensi (s)	TPS	Recall (%)	Presisi (%)	F1 (%)
S1 · Llama 3	569 ± 39	71,94 ± 15,39	9,05 ± 0,06	81,54 ± 11,67	82,90 ± 5,43	81,90 ± 7,39
S1 · Qwen 2.5	761 ± 63	92,37 ± 6,97	9,10 ± 0,10	72,31 ± 6,88	82,52 ± 5,80	76,98 ± 5,88
S1 · Mistral	790 ± 38	100,01 ± 15,25	8,86 ± 0,13	78,85 ± 11,54	78,92 ± 6,73	78,64 ± 7,91
S2 · Llama 3	520 ± 67	64,25 ± 19,30	9,31 ± 0,17	68,34 ± 3,73	100,00 ± 0,00	81,14 ± 2,56
S2 · Qwen 2.5	649 ± 97	85,33 ± 19,38	8,58 ± 0,38	73,33 ± 16,03	86,06 ± 11,72	78,78 ± 13,40
S2 · Mistral	776 ± 107	96,62 ± 13,36	8,87 ± 0,22	60,00 ± 16,03	87,98 ± 13,68	69,21 ± 8,44
S3 · Llama 3	476 ± 76	59,06 ± 12,70	9,29 ± 0,19	92,00 ± 4,47	100,00 ± 0,00	95,79 ± 2,35
S3 · Qwen 2.5	734 ± 132	86,31 ± 15,25	9,30 ± 0,11	92,00 ± 8,37	89,21 ± 7,23	90,15 ± 3,46

Skenario · Model	Token keluaran	Latensi (s)	TPS	Recall (%)	Presisi (%)	F1 (%)
S3 · Mistral	666 ± 122	81,65 ± 6,01	9,01 ± 0,37	92,00 ± 8,37	96,00 ± 5,48	93,78 ± 5,70

Struktur narasi. Penilaian rubrik dua penilai independen (Tabel 7) menempatkan Qwen 2.5 pada skor komposit tertinggi (4,60), diikuti Llama 3 (3,78) dan Mistral (3,51), dengan kesepakatan antar penilai substansial hingga hampir sempurna (κ_w 0,72–0,84, rerata 0,78).

Tabel 7. Skor Rubrik Penilai Independen dan Kesepakatan Antar Penilai (κ_w)

Dimensi rubrik	**Qwen 2.5**	**Llama 3**	**Mistral**	**κ_w**
D1 Struktur dan tata kelola informasi	4,78 ± 0,26	3,89 ± 0,42	3,39 ± 0,49	0,84
D2 Kelengkapan analisis dan diagnosis	4,44 ± 0,39	3,78 ± 0,44	3,44 ± 0,46	0,76
D3 Relevansi dan eksekusi rekomendasi	4,56 ± 0,39	3,50 ± 0,50	3,67 ± 0,43	0,79
D4 Keterbacaan dan eksekutif	4,61 ± 0,33	3,94 ± 0,46	3,56 ± 0,53	0,72
Komposit	4,60 ± 0,28	3,78 ± 0,41	3,51 ± 0,43	0,78

Pembahasan

Efisiensi komputasi

Throughput ketiga model relatif sama (selisih ≤ 3,3%) karena ukuran parameter, kuantisasi 4-bit, dan CPU-nya serupa, sehingga kecepatan inferensi lokal lebih dibatasi pembacaan bobot dari memori daripada arsitektur. Tabel 5 menunjukkan bahwa selisih latensi total terutama berasal dari panjang keluaran: Llama 3 rata-rata 521,73 token, sedangkan Qwen 2.5 (714,80) dan Mistral (740,93) masing-masing 37,0% dan 42,0% lebih panjang, sementara evaluasi prompt hanya sekitar 0,11 detik. Pada keluaran seragam 500 token, latensi ketiganya nyaris sama (54,34, 55,73, 56,16 detik). Dengan 65–92 detik per narasi, pembangkitan sebaiknya asinkron dengan hasil disimpan. Selisih peak RAM kurang dari 1 GB dan diukur pada tingkat sistem, sehingga hanya menunjukkan kelayakan pada server 16 GB (sisa sekitar 3–4 GB), bukan kebutuhan memori model.

Kualitas teks dan integritas angka

Perbedaan nilai rata-rata ROUGE dan *recall* numerik antar-model relatif kecil jika dibandingkan dengan simpangan baku hasil pengujian. Sebagai contoh, perbedaan rata-rata *recall* numerik antar-model hanya sekitar 1,4–3,8 poin persentase, sedangkan simpangan bakunya berada pada kisaran 12,2–18,2 poin. Oleh karena itu, tanpa pengujian statistik lebih lanjut dan jumlah sampel yang lebih besar, perbedaan nilai antara Llama 3, Qwen 2.5, dan Mistral belum dapat dianggap sebagai bukti keunggulan yang pasti.

Nilai BLEU-4 yang rendah (0,016–0,056 per skenario) tidak otomatis berarti narasi berkualitas rendah: BLEU menilai kesamaan susunan kata, sedangkan LLM dapat memparafrase dengan makna yang sama [3], dan satu teks acuan per skenario serta teks pendek juga menekan skor [4], [15]. Karena itu BLEU-4 (Qwen 2.5 0,052, Llama 3 0,041) tidak boleh menjadi satu-satunya dasar pemilihan model, dan evaluasi perlu dilengkapi metrik semantik seperti BERTScore [21] atau penilaian LLM [13], [23].

Recall numerik 76,81–80,62% berarti sekitar seperlima angka wajib belum tersampaikan. Pola ini bergantung pada skenario (Tabel 6): pada skenario krisis *recall* Mistral turun ke 60,00% karena angka SLA dan NPS tidak disebut, sedangkan pada skenario 3 ketiga model sama-sama 92,00%. Llama 3 unggul pada presisi (94,30%, 100% pada skenario 2 dan 3) dan F1 (86,28%), presisi Qwen 2.5 (85,93%) lebih rendah terutama karena ekstrapolasi sendiri, misalnya proyeksi kerugian kuartalan, yang dihitung sebagai FP.

Presisi dan F1 hanya menilai angka yang dapat diekstraksi dan dicocokkan menurut aturan, ekstrapolasi yang valid tetap dihitung FP. Karena itu, sistem produksi tetap memerlukan pemeriksa numerik otomatis terhadap data sumber [10].

Implikasi pemilihan model dan arsitektur

Ketiga model menunjukkan *trade-off* yang berbeda. Llama 3 memiliki *throughput*, *recall*, presisi, dan F1 numerik tertinggi dengan latensi total terendah (narasi lebih pendek), sehingga sesuai bila ketepatan angka dan kecepatan diutamakan. Qwen 2.5 menghasilkan narasi paling terstruktur dan mendalam (skor rubrik tertinggi) tetapi memakai memori lebih besar. Mistral memakai memori paling rendah dan narasinya ringkas, tetapi latensinya lebih tinggi dan *recall*-nya terendah pada skenario krisis.

Hasil ini masih berdasarkan pengujian terbatas dan penilaian kualitatif hanya mencakup sembilan narasi yang dinilai dua penilai, sehingga tidak dapat dianggap sebagai keunggulan mutlak suatu model.

Keterbatasan

Keterbatasan penelitian ini adalah: (1) data sintetis dengan tiga skenario dan satu teks acuan per skenario, (2) kuantisasi berbeda (Q4_0 untuk Llama 3, Q4_K_M untuk Qwen 2.5 dan Mistral), sehingga perbedaan kualitas dan memori belum sepenuhnya berasal dari arsitektur, (3) memori diukur pada tingkat sistem, (4) penilaian rubrik hanya melibatkan dua penilai pada sembilan narasi terpilih sehingga bersifat eksploratif, dan lima replikasi pada prompt yang sama bukan observasi bisnis independen, serta (5) satu rancangan *zero-shot prompt*, satu bahasa, satu perangkat berbasis CPU, replikasi terbatas, dan tanpa uji signifikansi statistik.

Penelitian selanjutnya dapat menggunakan data bisnis nyata, kuantisasi yang seragam, lebih banyak penilai independen dengan uji statistik, metrik semantik, model yang lebih besar, serta pendekatan *retrieval* dan *function calling* untuk mengambil angka langsung dari sumber data.

Kesimpulan

Penelitian ini menunjukkan bahwa LLM sumber terbuka berukuran kecil hingga menengah dapat digunakan untuk menghasilkan narasi wawasan bisnis secara otomatis pada dashboard Business Intelligence (BI) dan dijalankan secara lokal. Llama 3 menunjukkan presisi numerik tertinggi (94,30%) dan waktu pemrosesan total terendah, Qwen 2.5 menghasilkan skor rubrik tertinggi dari dua penilai independen (4,60 dari 5,00), sedangkan Mistral membutuhkan memori yang lebih rendah.

Kualitas narasi tidak cukup dinilai dari kesamaan kata, evaluasi perlu mempertimbangkan ketepatan informasi dan kesesuaian makna. Kinerja penerapan nyata dapat ditingkatkan melalui *caching*, *temperature* rendah, proses *asynchronous*, dan validasi angka, sedangkan penelitian selanjutnya dapat memakai data bisnis nyata, lebih banyak penilai, kuantisasi seragam, dan pendekatan *retrieval*.

Daftar Pustaka

[1] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[2] N. Carlini *et al.*, “Extracting training data from large language models,” in *Proc. 30th USENIX Security Symp.*, 2021, pp. 2633–2650.

[3] A. Celikyilmaz, E. Clark, and J. Gao, “Evaluation of text generation: A survey,” *arXiv preprint arXiv:2006.14799*, 2020.

[4] B. Chen and C. Cherry, “A systematic comparison of smoothing techniques for sentence-level BLEU,” in *Proc. 9th Workshop on Statistical Machine Translation*, Baltimore, MD, USA, 2014, pp. 362–367.

[5] H. Chen, R. H. L. Chiang, and V. C. Storey, “Business intelligence and analytics: From big data to big impact,” *MIS Quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[6] A. Dubey *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[7] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Official Journal of the European Union*, vol. L 119, pp. 1–88, 2016.

[8] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," in *Proc. 11th Int. Conf. Learning Representations (ICLR)*, 2023.

[9] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, 2018.

[10] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, 2023.

[11] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[13] Y. Liu *et al.*, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023.

[14] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[16] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[17] Republik Indonesia, *Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*, Jakarta, Indonesia, 2022.

[18] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[19] S. Wu *et al.*, "BloombergGPT: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[20] A. Yang *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. 8th Int. Conf. Learning Representations (ICLR)*, 2020.

[22] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[23] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[24] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[25] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.