JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.7987

ANALISIS KOMPARATIF KINERJA MINICPM-V2.6 SEBAGAI LLM MULTIMODAL DALAM VISUAL QUESTION ANSWERING PADA STRUK PEMBELIAN DIGITAL

Richo

Alumni Politeknik Perkapalan Negeri Surabaya; Jalan Teknik Kimia ITS Surabaya; (031) 5947186

Keywords:
LLM Multimodal;
MiniCPM-v2.6;
Struk Digital;
VQA.

Corespondent Email:
richo@student.ppns.ac.id

Abstrak. Meningkatnya volume transaksi digital dan kebutuhan otomatisasi pemrosesan dokumen, terutama dokumen semi-struktural seperti struk pembelian, maka diperlukan sistem cerdas yang mampu mengekstraksi informasi penting secara otomatis dan efisien. Namun, dokumen semacam ini umumnya memiliki format visual yang tidak konsisten, informasi numerik yang kompleks, dan tata letak tidak terstruktur, sehingga menimbulkan tantangan besar dalam proses ekstraksi informasi berbasis *Optical Character Recognition* (OCR) konvensional. Untuk menjawab tantangan tersebut, penelitian ini mengembangkan dan mengevaluasi sistem *Visual Question Answering* (VQA) berbasis *Large Language Model* (LLM) multimodal untuk mendeteksi dan memahami isi struk pembelian secara menyeluruh. Beberapa model VQA mutakhir seperti MiniCPM-v2.6, LLaMA-3, DeepSeek-VL2, LLaVA, dan BLIP-2 diuji menggunakan prompt engine multifungsi yang dirancang secara sistematis. Evaluasi dilakukan menggunakan metrik *BERT Cosine Accuracy* (BCA) untuk mengukur kesesuaian semantik antara jawaban model dan jawaban aktual, serta waktu inferensi sebagai indikator efisiensi eksekusi. Hasil menunjukkan bahwa MiniCPM-v2.6 unggul dengan rata-rata BCA sebesar 97,68% dan waktu eksekusi tercepat sekitar 5,51 menit. Dengan keunggulan ini, MiniCPM-v2.6 direkomendasikan sebagai model yang paling efisien dan akurat untuk sistem VQA berbasis dokumen semi-struktural, khususnya untuk implementasi dalam perangkat edge atau sistem kasir cerdas.

Abstract. *With the increasing volume of digital transactions and the growing demand for automated document processing—particularly for semi-structured documents such as purchase receipts—there is a critical need for intelligent systems capable of extracting essential information automatically and efficiently. However, such documents often exhibit inconsistent visual formats, complex numerical content, and unstructured layouts, posing significant challenges for conventional Optical Character Recognition (OCR)-based methods. To address these challenges, this study develops and evaluates a Visual Question Answering (VQA) system based on multimodal Large Language Models (LLMs) to detect and comprehend the contents of purchase receipts comprehensively. Several state-of-the-art VQA models, including MiniCPM-v2.6, LLaMA-3, DeepSeek-VL2, LLaVA, and BLIP-2, were tested using a systematically designed multifunctional prompt engine. The evaluation employed the BERT Cosine Accuracy (BCA) metric to measure the semantic similarity between the model-generated answers and the actual answers, as well as inference time as an indicator of computational efficiency. The results demonstrate that MiniCPM-v2.6 outperformed other models, achieving an average BCA of 97.68% and the fastest inference time of approximately 5.51*

minutes per execution. These findings highlight MiniCPM-v2.6 as the most accurate and efficient model for VQA tasks on semi-structured documents, making it particularly suitable for deployment in edge devices or smart cashier systems.

1. PENDAHULUAN

Dalam era transformasi digital yang semakin berkembang, proses otomasi menjadi kebutuhan mendesak di berbagai sektor, termasuk dalam pengolahan data dokumen. Namun, pada kenyataannya, masih banyak perusahaan yang melakukan proses pemindahan data dari dokumen fisik seperti struk belanja atau nota pembelian secara manual [1]. Praktik ini umumnya melibatkan tenaga admin yang bertugas untuk melakukan input nilai-nilai penting secara satu per satu ke dalam sistem. Selain menguras waktu dan tenaga, proses manual ini juga rentan terhadap kesalahan manusia (*human error*) yang dapat berdampak pada akurasi data serta pengambilan keputusan. Sebagai solusi atas permasalahan ini, dibutuhkan pendekatan berbasis teknologi yang mampu melakukan ekstraksi informasi secara otomatis, cepat, dan akurat dari dokumen bergambar seperti struk belanja.

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*/AI) telah membawa dampak signifikan terhadap cara manusia memproses dan memahami informasi visual, terutama melalui pendekatan *Visual Question Answering* (VQA). VQA merupakan bidang riset interdisipliner yang menggabungkan visi komputer (*computer vision*) dan pemrosesan bahasa alami (*natural language processing*) untuk menjawab pertanyaan berdasarkan gambar yang diberikan [2]. Dalam konteks praktis, VQA dapat digunakan untuk membantu mengekstraksi dan memahami dokumen gambar seperti struk pembelian.

Struk pembelian merupakan jenis dokumen yang umumnya dihasilkan secara digital oleh sistem kasir, dan memiliki komponen penting seperti tanggal transaksi, daftar barang, harga satuan, total belanja, diskon, serta nilai kembalian, yang biasanya tercetak dalam format tidak teratur dan terkadang dengan kualitas visual yang buruk akibat proses pemindaian atau pengambilan gambar. Struk pembelian menjadi salah satu objek visual yang menantang dalam VQA karena mengandung berbagai elemen teks tidak beraturan, variasi tata letak, serta kualitas citra yang sering kali rendah akibat hasil pemindaian atau pengambilan gambar yang tidak sempurna [3][4]. Oleh karena itu, diperlukan model multimodal yang mampu mengekstrak dan mengintegrasikan informasi dari dua modalitas utama, yakni teks dan gambar, secara efektif. Salah satu pendekatan yang semakin berkembang adalah pemanfaatan model multimodal transformer, seperti MiniCPM, yang didesain untuk memahami hubungan antara elemen visual dan linguistik dalam satu arsitektur terpadu [5].

MiniCPM merupakan model ringan turunan dari arsitektur CPM (*Chinese Pre-trained Models*) yang telah mengalami penyederhanaan agar tetap efisien namun lebih kompeten dalam tugas-tugas multimodal, termasuk VQA. Model ini memanfaatkan teknik alignment antara representasi visual dan representasi linguistik untuk menjawab pertanyaan berbasis citra. Dalam studi terkini, model-model seperti MiniCPM telah diuji pada berbagai benchmark dataset seperti VQA v2.0, GQA, dan OK-VQA, menunjukkan performa yang cukup menjanjikan dibandingkan dengan model besar seperti Llama atau DeepSeek [6]. Meskipun demikian, penerapan MiniCPM untuk konteks VQA pada dokumen seperti struk pembelian masih tergolong minim dieksplorasi. Mayoritas studi VQA berfokus pada citra umum atau *natural image*, bukan dokumen OCR-based yang memiliki struktur unik dan memerlukan pengenalan teks (OCR) terlebih dahulu [7]. Oleh karena itu, analisis komparatif kinerja MiniCPM dalam domain ini menjadi penting untuk memahami sejauh mana kemampuannya menangani konteks visual-linguistik yang kompleks dan spesifik, serta dibandingkan dengan model lain dalam segmen yang sama [8]. Pada penelitian ini peneliti membandingkan metode MiniCPM-v 2.6 dengan model Multimodal lainnya seperti LlaVa, Llama-3, DeepSeek-VL2, dan BLIP-2.

Penelitian ini bertujuan untuk menganalisis kinerja MiniCPM sebagai model multimodal dalam tugas VQA khusus pada dokumen struk

pembelian, serta membandingkannya dengan beberapa model sejenis lainnya. Beberapa aspek evaluasi yang digunakan meliputi ketepatan jawaban (*accuracy*), kecepatan inferensi, dan efisiensi sumber daya komputasi. Dengan menggunakan dataset yang terdiri dari kumpulan struk pembelian nyata dan pertanyaan terstruktur, penelitian ini diharapkan dapat memberikan kontribusi nyata terhadap pengembangan solusi otomatisasi dokumen dalam dunia ritel dan akuntansi [9].

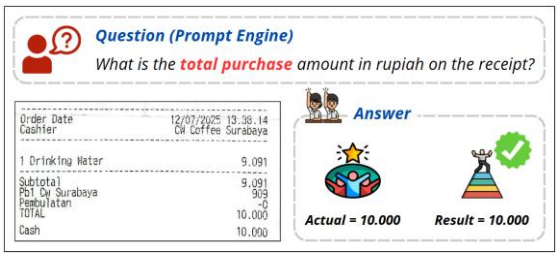
Hasil dari penelitian ini tidak hanya menunjukkan posisi relatif MiniCPM dibandingkan model lainnya, tetapi juga mengungkap tantangan-tantangan teknis dalam implementasi VQA untuk dokumen tekstual yang kompleks. Selain itu, temuan dari penelitian ini dapat menjadi dasar bagi pengembangan sistem cerdas berbasis OCR+VQA yang mampu menjawab pertanyaan seputar transaksi pembelian, validasi data, dan deteksi anomali pembelanjaan secara otomatis [3]. Dengan meningkatnya kebutuhan akan efisiensi analisis data transaksi dalam dunia bisnis, kajian ini diharapkan dapat menjadi referensi dalam implementasi sistem visual-tekstual berbasis AI.

2. TINJAUAN PUSTAKA

2.1 Visual Question Answering (VQA)

Visual Question Answering (VQA) merupakan salah satu bidang riset dalam kecerdasan buatan yang mengintegrasikan visi komputer dan pemrosesan bahasa alami untuk memungkinkan sistem menjawab pertanyaan berbasis gambar [10]. Model VQA umumnya terdiri dari dua komponen utama, ekstraktor fitur visual dan pengolah bahasa alami. Model seperti VQA v2.0 dan GQA menjadi benchmark umum dalam menilai performa algoritma VQA. Penelitian awal yang dilakukan oleh Pratomo et al. [11] menunjukkan bagaimana pendekatan deep learning berbasis CNN dan OCR digunakan untuk ekstraksi informasi dari gambar visual. Meskipun pendekatan ini cukup efektif dalam konteks visual umum, performa model-model tersebut cenderung menurun ketika diterapkan pada dokumen teks seperti struk atau laporan. Dokumen-dokumen tersebut memiliki struktur visual yang lebih kompleks karena mengandung kombinasi elemen tekstual dan spasial seperti tabel, label harga, logo, hingga format penulisan khusus yang tidak seragam. Struktur khas pada dokumen seperti struk pembelian menghadirkan tantangan tersendiri, sebab proses ekstraksi informasi tidak hanya melibatkan pengenalan karakter optik (OCR), tetapi juga pemahaman konteks dari informasi yang tersebar secara spasial. Dengan demikian, dibutuhkan pendekatan multimodal yang tidak hanya mampu memahami teks, tetapi juga konteks visual secara menyeluruh. Dalam hal ini, penggunaan *Large Language Models* (LLM) berbasis multimodal menjadi sangat relevan.

Penelitian selanjutnya oleh Qardafil et al. [10], yang berjudul "Peningkatan Efektivitas Kualitas Chatbot Jala Tech Melalui Implementasi Chat GPT, Auto-GPT, dan Langchain", menunjukkan pemanfaatan LLM berbasis teks untuk meningkatkan kualitas sistem *chatbot*. Namun, pendekatan tersebut masih terbatas pada input berbasis teks saja. Hal ini menjadi tantangan tersendiri ketika sistem dihadapkan dengan input dalam bentuk gambar dokumen, seperti struk belanja, yang memerlukan pemrosesan visual sebelum dilakukan interpretasi konteks. Oleh karena itu, dibutuhkan pengembangan sistem yang mampu menangani input multimodal secara langsung yakni gambar dan teks dalam satu kesatuan proses. Pendekatan berbasis LLM Multimodal dikembangkan untuk menjawab pertanyaan berbasis dokumen digital yang sudah dipindai, termasuk laporan, invoice, dan formulir. Dalam konteks VQA (*Visual Question Answering*), LLM multimodal memainkan peran krusial karena model ini tidak hanya memahami isi pertanyaan (teks), tetapi juga mampu memahami dan menafsirkan konten visual (gambar) untuk menghasilkan jawaban yang kontekstual dan logis. Kombinasi dua modalitas ini (visual dan linguistik) menjadikan LLM multimodal sebagai solusi yang sangat efektif untuk tugas-tugas berbasis pemahaman visual-linguistik. Pada Gambar 1 berikut ini merupakan kinerja LLM multimodal.



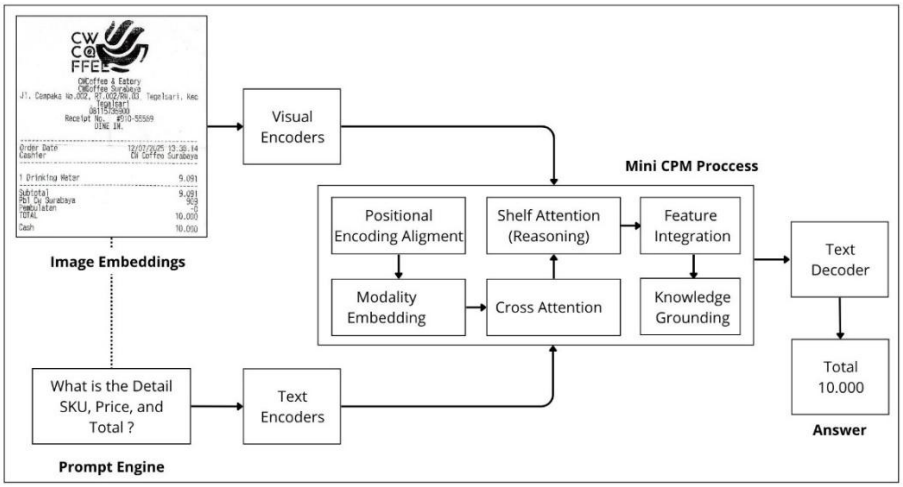
Gambar 1. Kinerja LLM VQA Secara Sederhana

LLM multimodal menerima dua jenis input secara simultan, yaitu *input* visual berupa gambar dokumen dan *input* teks berupa pertanyaan atau *prompt engine* yang telah dirancang sebelumnya. *Input* visual berfungsi untuk memahami isi visual dari dokumen, termasuk teks, angka, serta tata letak (*layout*) yang membentuk struktur informasi dalam gambar. Sementara itu, *input teks* berperan dalam memberikan konteks pertanyaan yang diajukan, seperti makna frasa "*total purchase amount*" dan pengenalan terhadap satuan mata uang seperti rupiah. Setelah kedua input diproses, model melakukan tahapan *reasoning* atau penalaran semantik, yaitu mencocokkan informasi yang tersedia dalam gambar dengan maksud dari pertanyaan. Pada proses ini, model mengenali bahwa angka 10.000 merupakan jawaban yang relevan karena sesuai dengan konteks pertanyaan dan isi dokumen yang telah dipahami secara multimodal.

2.2 MiniCPM Model

Dalam beberapa tahun terakhir, telah dikembangkan berbagai model *Large Language Model* (LLM) multimodal yang mampu memahami dan mengintegrasikan informasi visual serta teks secara simultan untuk menyelesaikan tugas-tugas seperti *Visual Question Answering* (VQA). Beberapa model yang menonjol antara lain LLaVA, LLaMA-3, DeepSeek-VL2, BLIP-2, serta MiniCPM-v2.6. Setiap model memiliki pendekatan arsitektur dan strategi pelatihan yang berbeda-beda, yang berdampak pada performa, efisiensi, serta tingkat akurasi dalam memahami dokumen semi-struktural seperti struk pembelian.

Dalam penelitian ini, penulis memilih untuk menggunakan model MiniCPM-v2.6 karena model ini dirancang dengan arsitektur ringan dan efisien, sehingga lebih cocok untuk skenario *real-time* dan implementasi pada perangkat dengan sumber daya terbatas, seperti sistem kasir berbasis AI atau aplikasi mobile. MiniCPM-v2.6 tidak hanya unggul dari sisi ukuran parameter yang relatif kecil, tetapi juga menunjukkan kinerja reasoning multimodal yang kompetitif dibandingkan model-model LLM multimodal lainnya. Dengan tetap mempertahankan ketajaman pemahaman visual-linguistik, MiniCPM-v2.6 mampu memberikan hasil prediksi yang akurat, bahkan dalam dokumen dengan struktur visual kompleks atau kualitas gambar yang rendah. Dengan mempertimbangkan efisiensi komputasi, kesesuaian terhadap jenis dokumen yang digunakan (yakni struk pembelian), serta kemudahan integrasi ke dalam sistem ringan, MiniCPM-v2.6 menjadi pilihan yang optimal untuk tugas *Visual Question Answering* dalam konteks penelitian ini.



Gambar 2. Sistematika Kinerja Model MiniCPM

Berdasarkan gambar diagram arsitektur di atas, Proses dimulai dari dua jenis *input* utama: *image embeddings* dan *text embeddings*. Gambar struk terlebih dahulu diproses oleh *Visual Encoders* untuk mengekstrak fitur visual, sedangkan pertanyaan atau prompt dalam bentuk teks, seperti "*What is the Detail SKU, Price, and Total?*", diproses oleh *Text Encoders* untuk menghasilkan representasi teks. Kedua jenis representasi ini kemudian dikombinasikan dalam tahap *Cross Attention*, yang memungkinkan model untuk menyelaraskan informasi dari dua modalitas berbeda.

Setelah proses atensi silang, MiniCPM menjalankan sejumlah modul lanjutan seperti *Positional Encoding Alignment, Modality Embedding, Shelf Attention (Reasoning), Feature Integration, dan Knowledge Grounding*. Modul-modul ini bekerja secara sinergis untuk menyatukan informasi spasial (misalnya posisi angka pada struk), semantik (makna teks), serta logika penalaran untuk memahami struktur dan konteks informasi. Hasil dari proses ini kemudian diteruskan ke *Text Decoder* yang bertugas menghasilkan jawaban dalam bentuk teks, seperti total pembelian (misalnya "Total 10.000"). Dengan cara kerja seperti ini, MiniCPM mampu memahami dan menjawab pertanyaan spesifik dari dokumen kompleks berbasis visual, serta mengurangi ketergantungan terhadap input manual atau proses OCR tradisional.

2.3 Perancangan VQA dengan MiniCPM

Perancangan sistem *Visual Question Answering* (VQA) dalam penelitian ini difokuskan pada objek berupa dokumen struk pembelian digital yang berbahasa Indonesia maupun Inggris. Sistem ini dirancang untuk mengevaluasi kemampuan pemahaman dokumen *visual-teks* oleh model *Large Language Model* (LLM) multimodal melalui pendekatan komparatif [12]. Beberapa model yang dibandingkan dalam eksperimen ini mencakup LLaVA, LLaMA-3, DeepSeek-VL2, BLIP-2, dan MiniCPM-v2.6. Setiap model diuji dalam kondisi variatif yang merepresentasikan kompleksitas struktur struk pembelian, seperti keberagaman tata letak, jenis font, posisi informasi, dan bahasa.

Tujuan dari pendekatan komparatif ini adalah untuk menilai sejauh mana masing-masing model mampu memahami konteks visual dan tekstual secara terpadu, serta memberikan jawaban yang relevan dan akurat terhadap pertanyaan spesifik yang diajukan terhadap isi struk. Evaluasi dilakukan secara menyeluruh berdasarkan kualitas jawaban, ketepatan informasi, serta efisiensi pemrosesan [13]. Dengan demikian, proses ini tidak hanya bertujuan untuk mengetahui performa individu dari masing-masing model, tetapi juga untuk mengidentifikasi model paling optimal yang dapat diimplementasikan secara nyata dalam sistem otomasi ekstraksi data dokumen berbasis VQA.

3. METODE PENELITIAN

3.1 Analisis dan Identifikasi Masalah

Tahap awal dalam penelitian ini diawali dengan proses analisis kebutuhan, yang bertujuan untuk memahami permasalahan utama yang dihadapi dalam proses ekstraksi informasi dan jawaban atas pertanyaan berbasis visual dari dokumen semi-struktural seperti struk pembelian. Dalam tahap ini dilakukan identifikasi terhadap kesenjangan yang ada dalam teknologi pengolahan dokumen berbasis OCR dan *Visual Question Answering* (VQA), serta potensi keterbatasan model-model existing. Dengan memahami tantangan teknis dan kebutuhan praktis tersebut, peneliti dapat merumuskan fokus penelitian secara spesifik dan terarah.

3.2 Pengumpulan Data (Data Collecting)

Setelah masalah diidentifikasi, langkah selanjutnya adalah melakukan pengumpulan data berupa dataset gambar struk pembelian yang representatif. Dataset ini diperoleh dari berbagai sumber seperti hasil pemindaian struk asli, dokumentasi digital toko ritel, serta sumber publik jika tersedia. Setiap data dilengkapi dengan anotasi pertanyaan dan jawaban yang relevan, sehingga dapat digunakan sebagai basis untuk pelatihan dan evaluasi model VQA. Kualitas dan keberagaman data dijaga agar mencerminkan berbagai variasi format dan kondisi visual struk pembelian di dunia nyata.

3.3 Penelitian dan Komparasi Metode

Pada tahap ini dilakukan kajian literatur dan eksplorasi berbagai metode model VQA

multimodal yang relevan, seperti MiniCPM-v 2.6, LlaVa, Llama-3, DeepSeek-VL2, serta BLIP-2. Model - model tersebut mampu diimplementasikan dan diuji dalam konteks spesifik dokumen struk pembelian. Proses ini mencakup eksperimen awal untuk menilai performa masing-masing model dalam hal akurasi, kecepatan inferensi, serta kemampuan memahami informasi numerik dan tekstual dari gambar dokumen. Tujuan dari tahap ini adalah memperoleh pemahaman komparatif terhadap kekuatan dan kelemahan masing-masing metode dalam domain aplikasi yang diteliti.

3.4 Penyesuaian dan Desain Prompt Engine

Langkah selanjutnya adalah melakukan penyesuaian terhadap *prompt* atau *input query* yang diberikan ke model sebagai question, agar sesuai dengan struktur pertanyaan yang umum diajukan pada struk pembelian. Dalam konteks model berbasis LLM dan multimodal, penyusunan *prompt* yang tepat sangat menentukan kualitas jawaban yang dihasilkan. Oleh karena itu, dilakukan eksperimen dengan berbagai desain *prompt engine* untuk mengoptimalkan pemetaan antara pertanyaan, isi gambar, dan output jawaban. Tahap ini juga mempertimbangkan strategi prompt-tuning yang mendukung pemrosesan semi-struktural [14].

3.5 Evaluasi Metode (Best Model Selection)

Setelah dilakukan pengujian dan penyetelan prompt, seluruh model yang diuji dievaluasi menggunakan metrik kuantitatif seperti *accuracy, exact match score,* serta metrik efisiensi seperti waktu inferensi dan penggunaan sumber daya. Hasil evaluasi digunakan untuk menentukan model terbaik yang paling sesuai untuk tugas VQA pada struk pembelian. Analisis komparatif dilakukan secara objektif untuk memastikan bahwa model terpilih memiliki performa yang konsisten dan unggul pada berbagai skenario data.

3.6 Tunning dan Efisiensi Komputasi

Tahap akhir dari penelitian ini adalah melakukan penyetelan parameter (parameter *tuning*) pada model terpilih untuk meningkatkan akurasi prediksi sekaligus mengefisienkan waktu pemrosesan. Proses ini mencakup pengaturan hyperparameter seperti *learning rate, batch size*, serta konfigurasi pemrosesan visual. Selain itu, dilakukan optimasi terhadap *pipeline* inferensi agar model dapat berjalan lebih cepat, ringan, dan efisien, terutama untuk kebutuhan implementasi di lingkungan terbatas sumber daya seperti perangkat edge atau sistem kasir cerdas [15].

4. HASIL DAN PEMBAHASAN

4.1 Evaluasi Komparasi Metode

Pada bagian ini dilakukan evaluasi komparatif terhadap sejumlah metode LLM multimodal yang telah dipilih sebelumnya. Pengujian difokuskan pada kemampuan setiap model dalam merespons pertanyaan berbasis visual melalui berbagai variasi struktur *prompt engine* yang dirancang secara sistematis.

Tujuan utama dari pengujian ini adalah untuk memperoleh pemahaman yang mendalam mengenai performa relatif dari tiap-tiap model dalam konteks aplikasi dunia nyata, di mana informasi visual tidak selalu tersaji dalam format yang sepenuhnya terstruktur. Dengan pendekatan ini, diharapkan dapat diidentifikasi kekuatan dan keterbatasan masing-masing model dalam menginterpretasi serta mengekstraksi informasi penting dari gambar dokumen.

Sebagai ilustrasi, Gambar 3 menyajikan sejumlah potongan data visual yang digunakan dalam proses pengujian. Gambar-gambar tersebut merepresentasikan beragam jenis struk pembelian yang memiliki perbedaan dalam format, kualitas pencetakan, serta kompleksitas tata letak, sehingga memberikan tantangan tersendiri bagi model dalam memahami dan merespons input yang diberikan.

Gambar 3. Visual Struk Pembelian yang Diuji

Pengujian dilakukan dengan mempertimbangkan nilai *BERT Cosine Accuracy* (BCA), yaitu persentase kemiripan antara jawaban yang dihasilkan oleh masing-masing metode dengan jawaban aktual (*Correct Actual Answer*). Nilai BCA diperoleh melalui representasi vektor dari setiap teks menggunakan model *Sentence-BERT*, kemudian dihitung tingkat kemiripan nya menggunakan rumus cosine similarity antar dua vektor. Selain aspek akurasi semantik, pengujian juga mempertimbangkan rata-rata waktu eksekusi (*performance time*) dari masing-masing metode untuk mengevaluasi efisiensi komputasi dalam proses inferensi. Persamaan (1) berikut ini merupakan rumus perhitungan *BERT Cosine Accuracy* (BCA) yang digunakan dalam penelitian ini, sedangkan hasil komparatif antar model disajikan pada Tabel 1.

$$\text{Cosine similirity (M, N)} = \frac{\sum_{l=1}^{k} Mi.Ni}{\sqrt{\sum_{l=1}^{k} Mi^2 \; . \; \sum_{l=1}^{k} Ni^2}}$$

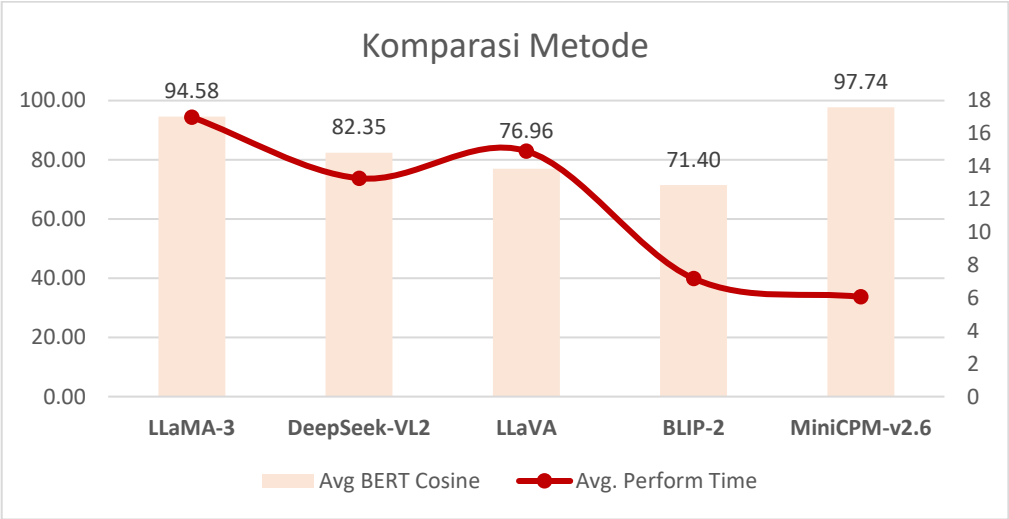
Persamaan *BERT Cosine Similarity* sebagaimana ditunjukkan pada Persamaan (1) digunakan untuk mengukur tingkat kemiripan semantik antara dua kalimat yang telah direpresentasikan dalam bentuk vektor melalui model BERT. Perhitungan ini didasarkan pada *cosine similarity*, yaitu rasio antara hasil perkalian *dot product* dua vektor (jawaban model dan jawaban aktual) terhadap hasil perkalian norma (*magnitudo*) dari masing-masing vektor tersebut.

Nilai yang dihasilkan berada dalam rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan kemiripan semantik yang tinggi, dan nilai mendekati 0 menunjukkan perbedaan makna yang signifikan. Dengan pendekatan ini, evaluasi performa model tidak hanya mempertimbangkan kesamaan kata secara literal, tetapi juga kedekatan makna dalam konteks bahasa alami dan maksud yang dituju, sehingga lebih representatif dalam menilai kualitas jawaban sistem VQA berbasis LLM multimodal.

Dengan mempertimbangkan kedekatan makna dalam konteks bahasa alami dan maksud pertanyaan yang diajukan, pendekatan ini memungkinkan penilaian yang lebih representatif terhadap kualitas jawaban yang dihasilkan oleh sistem *Visual Question Answering* (VQA) berbasis LLM multimodal. Hal ini penting karena dalam banyak kasus, jawaban yang berbeda secara tekstual dapat memiliki makna yang serupa, atau sebaliknya, jawaban yang tampak serupa secara kata dapat berbeda secara semantik. Oleh karena itu, penggunaan metrik semantik dalam evaluasi performa model tidak hanya meningkatkan akurasi penilaian, tetapi juga mencerminkan kemampuan model dalam memahami konteks

Tabel 1. Komparasi Berbagai LLM Multimodal Terhadap Respon VQA

Correct Actual Answer	Methods	% BERT Cosine (Acc.)			Avg BERT Cosine (Acc.) (%)	Avg. Perform Time (minutes)
		Image 1	Image 2	Image 3		
Prompt *: Pada tanggal berapa transaksi pembelian pada struk dilakukan?*						
Image 1 : 12/07/2025 Image 2 : 12/07/2025 Image 3 : 10 Jul 2025	LLaMA-3	95.85	94.35	95.38	95.19	14.53
	DeepSeek-VL2	89.66	87.48	83.56	86.90	9.67
	LLaVA	82.91	71.96	70.09	74.99	10.25
	BLIP-2	80.24	76.25	60.85	72.45	6.32
	MiniCPM-v2.6	97.89	97.21	96.55	97.22	5.02
Prompt *: Menu apa yang dibeli pada struk Pembelian?*						
Image 1 : 1 Drinking water Image 2 : Caffe Latte Dingin Image 3 : Basreng Koin Super Pedas dan Basreng Koin Ori Super	LLaMA-3	94.21	93.56	85.33	91.03	17.95
	DeepSeek-VL2	85.63	82.79	79.89	82.77	14.62
	LLaVA	81.75	80.55	79.52	80.77	15.23
	BLIP-2	70.12	75.16	79.47	74.92	6.97
	MiniCPM-v2.6	97.54	98.21	97.05	97.60	7.05
Prompt *: Calculate the total transaction amount in rupiah from the receipt image?*						
Image 1 : 10.000 Image 2 : 27.000 Image 3 : 128.300	LLaMA-3	98.98	96.52	97.10	97.53	18.51
	DeepSeek-VL2	85.56	79.16	67.42	77.38	15.54
	LLaVA	75.29	79.85	70.21	75.12	19.32
	BLIP-2	80.56	56.19	63.78	66.84	8.26
	MiniCPM-v2.6	97.51	98.95	98.77	98.41	8.15



Gambar 4 Komparasi Model Terhadap Nilai *BERT Consine* dan *Perform Time*

Berdasarkan hasil pengujian dari tiga skenario pertanyaan yang melibatkan tanggal transaksi, daftar item pembelian, dan total nilai transaksi, model MiniCPM-v2.6 menunjukkan performa yang paling unggul secara konsisten. Model ini mencatatkan akurasi semantik tertinggi pada seluruh jenis pertanyaan berdasarkan metrik BERT *Cosine Accuracy* (BCA), sekaligus menunjukkan efisiensi waktu eksekusi yang sangat baik. Kombinasi antara akurasi dan efisiensi tersebut menjadikan MiniCPM-v2.6 sangat sesuai untuk diterapkan pada sistem dengan keterbatasan sumber daya komputasi, seperti perangkat *edge* atau sistem kasir cerdas. Sementara itu, model LLaMA-3 memberikan hasil akurasi yang kompetitif, namun memiliki waktu pemrosesan yang relatif lama, sehingga kurang optimal dari sisi efisiensi.

Model DeepSeek-VL2 dan LLaVA menunjukkan performa sedang namun cenderung fluktuatif antar jenis pertanyaan, yang mengindikasikan keterbatasan dalam konsistensi pemrosesan informasi semi-struktural. Di sisi lain, BLIP-2 meskipun cukup cepat dalam eksekusi, memiliki tingkat akurasi yang rendah terutama dalam pertanyaan yang melibatkan elemen numerik dan tanggal,

sehingga kurang direkomendasikan untuk konteks ini. Oleh karena itu, dapat disimpulkan bahwa MiniCPM-v2.6 merupakan pilihan model paling seimbang dan efektif untuk mendukung sistem otomatisasi berbasis *Visual Question Answering* pada dokumen semi-struktural seperti struk pembelian.

Hal ini dibuktikan dengan perolehan nilai rata-rata *BERT Cosine Accuracy* (BCA) tertinggi sebesar 97,74%, yang menunjukkan tingkat kemiripan semantik jawaban model dengan jawaban aktual lebih unggul dibandingkan model lain yang diuji. Selain itu, model ini juga mencatat waktu rata-rata eksekusi tercepat, yaitu sekitar 6,06 menit, sehingga menunjukkan efisiensi komputasi yang lebih baik dibandingkan model-model lainnya. Kombinasi antara keakuratan semantik yang tinggi dan efisiensi waktu pemrosesan ini menjadikan model semakin baik, di mana kecepatan dan ketepatan respon merupakan faktor krusial dalam meningkatkan pengalaman pengguna dan efektivitas sistem secara keseluruhan.

4.2 Tuning Arsitektur Model MiniCPM-V2.6

Pada tahap ini dilakukan proses tuning terhadap arsitektur model MiniCPM-v2.6 guna mengoptimalkan performa dalam tugas ekstraksi informasi dari struk pembelian. Penyesuaian dilakukan pada parameter inferensi seperti *temperature, top_p, top_k, repetition_penalty, dan max_new_tokens* yang bertujuan untuk meningkatkan akurasi semantik jawaban serta efisiensi waktu eksekusi. Selain itu, penggunaan model versi int4 dipilih untuk mengurangi beban komputasi tanpa mengorbankan performa *predictive* secara signifikan. Dengan tuning parameter yang tepat, model mampu memberikan hasil keluaran yang lebih relevan dan stabil pada berbagai skenario prompt yang diuji.

Tabel 2. Tunning Parameter Arsitektur Model MiniCPM-v2.6

Komponen Arsitektur MiniCPM-v2.6	**Nilai/Konfigurasi**	**Deskripsi**
Model Architecture	MiniCPM For Casual LM	Arsitektur dasar untuk *generative language model*
Hidden Size	4096	Ukuran dimensi vektor per token (*per layer*)
Intermediated Size	1108	Ukuran *feedforward layer* dalam blok Transformer
Number of Attention Heads	32	Jumlah kepala dalam mekanisme *multi-head attention*
Number of Hidden Layers	32	Jumlah blok Transformer (*depth*)
Vocabulary Size	50032	Jumlah total token yang dikenali dalam tokenizer
Mas Position Embeddings	4096	Panjang maksimal urutan *input* token
Activation Function	Silu	Fungsi aktivasi dalam *feedforward* (*Smooth ReLU*)
Layer Normalization Epsilon	1e-5	Nilai *epsilon* untuk menghindari pembagian nol pada normalisasi
Torch Dtype	Int4	Model dikompresi menggunakan *INT4 quantization* untuk efisiensi inferensi

Berdasarkan Tabel 2 diatas, Model MiniCPM-V-2.6 yang digunakan telah dikompresi dalam representasi INT4 untuk mengurangi ukuran model dan mempercepat inferensi, tanpa kehilangan akurasi yang signifikan. Seluruh eksperimen dijalankan pada mode evaluasi (model.eval()) dan menggunakan GPU dengan dukungan CUDA.

4.3 Pengujian Sistem dengan MiniCPM-V2.6

Setelah proses tuning arsitektur dilakukan, tahap selanjutnya adalah melakukan pengujian sistem menggunakan model MiniCPM-v2.6 yang telah dikonfigurasi. Pengujian difokuskan pada kemampuan model dalam menjawab pertanyaan berbasis visual terhadap dokumen semi-struktural berupa struk pembelian. Sistem diuji dengan memberikan input berupa gambar struk dan prompt tertentu, kemudian dianalisis hasil keluaran model baik dari segi ketepatan jawaban (akurasi semantik) maupun efisiensi waktu proses inferensi. Evaluasi dilakukan dengan pendekatan berbasis *BERT Cosine Accuracy* (BCA) untuk mengukur tingkat kemiripan semantik antara jawaban model dengan jawaban aktual, serta pemantauan waktu eksekusi untuk mengukur kinerja sistem secara keseluruhan. Pengujian dilakukan dengan *question* (*prompt engine*) seperti yang tertera pada Gambar 5 berikut ini.

```
prompt = f"""
Lakukan ekstraksi informasi dari gambar struk pembelian berikut dan tampilkan hasilnya dalam format yang jelas dan terstruktur.
Informasi yang perlu diekstrak meliputi:

1. Identitas pemilik atau asal struk pembelian (nama usaha dan alamat jika tersedia).
2. Tanggal dan waktu transaksi pembelian (Order Date).
3. Daftar barang yang dibeli, meliputi: kuantitas, nama barang, harga per item, total harga per item, dan informasi PPN atau pajak lainnya (jika tercantum).
4. Total akhir dari seluruh transaksi pembelian dalam satuan rupiah.
"""
```

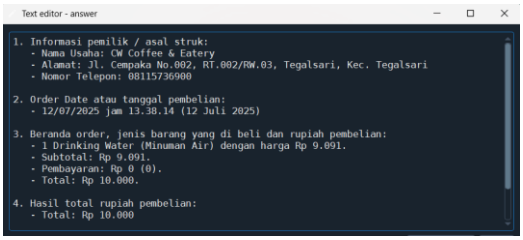
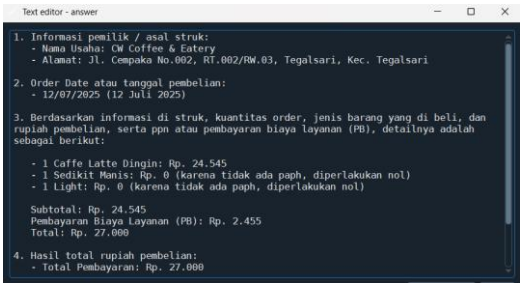
Gambar 5. Prompt Engine Multifunctional yang Digunakan

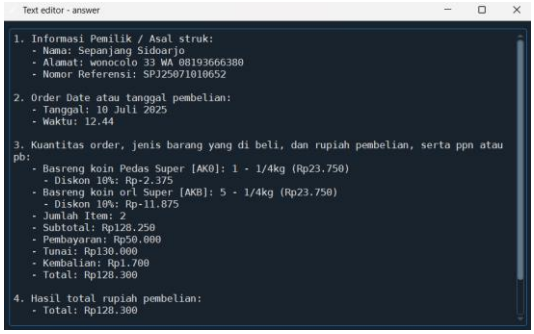
Prompt yang digunakan dalam pengujian ini dirancang dengan struktur yang sistematis dan terorganisir secara cermat untuk mengarahkan model dalam melakukan ekstraksi informasi kunci dari dokumen struk pembelian secara efektif. *Prompt* tersebut terdiri dari empat elemen utama yang disusun secara eksplisit dan berurutan, yaitu: identitas pemilik struk, tanggal terjadinya transaksi, rincian lengkap barang yang dibeli (meliputi kuantitas, nama barang, harga satuan, serta komponen pajak yang berlaku), dan terakhir adalah total nilai akhir dari pembelian tersebut.

Penyusunan *prompt* dengan pendekatan yang terperinci ini bertujuan untuk memaksimalkan kemampuan model dalam memahami dan menafsirkan informasi yang terdapat pada dokumen semi-struktural, sehingga dapat menghasilkan jawaban yang tidak hanya akurat secara literal, tetapi juga tepat secara semantik. Dengan cara ini, sistem *Visual Question Answering* (VQA) berbasis LLM multimodal dapat memberikan respons yang lebih relevan dan informatif dalam konteks pengolahan dokumen yang kompleks dan beragam formatnya, seperti struk pembelian.

Pendekatan terstruktur dalam penyusunan *prompt* ini juga diharapkan mampu mengurangi *ambiguities* dan kesalahan interpretasi yang sering muncul pada model ketika berhadapan dengan data yang memiliki variasi pola penyajian, sehingga secara keseluruhan meningkatkan kualitas serta konsistensi performa model dalam aplikasi nyata.

Tabel 3. Tunning Parameter Arsitektur Model MiniCPM-v2.6

Input Gambar	Hasil Ouput	Avg. BERT Cosine (%)	Avg. Perform Time (minutes)
		98.87	5.16
		95.61	5.33

		98.55	6.05
Persentase Rata-rata BERT Cosine Accuracy & Perform Time		97.68	5.51

Berdasarkan hasil pada Tabel 2 yang diperoleh dari pengujian terhadap tiga gambar input menggunakan prompt engine multifunctional yang telah dirancang secara sistematis, sistem *Visual Question Answering* (VQA) berbasis model MiniCPM-v2.6 menunjukkan performa yang sangat optimal. Model ini mampu memberikan kontribusi akurasi yang sangat baik dalam mengekstraksi informasi dari dokumen struk pembelian. Dari hasil evaluasi yang dilakukan, diperoleh nilai rata-rata *BERT Cosine Accuracy* (BCA) sebesar 97,68%, yang mencerminkan tingkat kemiripan semantik yang sangat tinggi antara jawaban model dan jawaban aktual. Selain keunggulan akurasi, model ini juga menunjukkan efisiensi dari sisi waktu, dengan rata-rata waktu eksekusi untuk satu kali inferensi hanya sekitar 5,51 menit, menjadikannya sebagai salah satu model yang paling responsif dan efisien dibandingkan model-model lain yang diuji dalam studi ini. Kinerja yang konsisten dan stabil pada berbagai jenis pertanyaan menjadikannya model yang adaptif dalam konteks dokumen semi-struktural. Oleh karena itu, MiniCPM-v2.6 direkomendasikan sebagai pilihan utama untuk solusi VQA pada dokumen transaksi digital seperti struk pembelian.

5. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa model MiniCPM-v2.6 merupakan solusi yang efektif dan efisien untuk tugas *Visual Question Answering* (VQA) pada dokumen semi-struktural, khususnya struk pembelian. Melalui pengujian yang dilakukan terhadap beberapa skenario gambar input dan menggunakan prompt engine multifungsi yang telah dirancang, model ini secara konsisten menunjukkan performa terbaik dibandingkan dengan model LLM multimodal lainnya seperti LLaMA-3, DeepSeek-VL2, LLaVA, dan BLIP-2.

MiniCPM-v2.6 memperoleh rata-rata nilai *BERT Cosine Accuracy* (BCA) sebesar 97,68%, yang mencerminkan tingkat kemiripan semantik yang sangat tinggi terhadap jawaban aktual. Selain itu, model ini juga mencatat waktu eksekusi tercepat, dengan rata-rata waktu inferensi hanya sekitar 5,51 menit, menjadikannya unggul tidak hanya dalam aspek akurasi, tetapi juga efisiensi komputasi. Keunggulan ini diperkuat oleh penggunaan arsitektur yang telah dioptimasi serta strategi tuning parameter yang mendukung kinerja pada perangkat dengan keterbatasan sumber daya. Secara keseluruhan, MiniCPM-v2.6 terbukti sebagai model paling seimbang dari segi akurasi semantik, stabilitas jawaban, dan efisiensi waktu, sehingga sangat direkomendasikan untuk diimplementasikan dalam sistem otomatisasi berbasis dokumen seperti kasir cerdas, perangkat edge, atau sistem VQA dokumen digital lainnya.

DAFTAR PUSTAKA

[1] Y. Astuti and K. K. Wicaksana, "Rancang Bangun Sistem Pemindaian Struk Belanja untuk Mendapatkan Rincian Belanja," Semin. Nas. Teknol. Inf. dan Multimed., vol. 6, no. 1, pp. 37–42, 2018.

[2] G. Lee and X. Zhai, "Realizing Visual Question Answering for Education: GPT-4V as a Multimodal AI," TechTrends, vol. 69, no. 2, pp. 271–287, 2025, doi: 10.1007/s11528-024-01035-z.

[3] L. S. Arifiyanto and F. Masya, "Analisa dan Perancangan Sistem Struk Digital Berbasis Android dan SMS Gateway," J. Sist. Inf. dan E-Bisnis, vol. 1, no. 6, pp. 214–222, 2019.

[4] G. Singh et al., "Efficiently Serving Large Multimodal Models Using EPD Disaggregation," 2024.

[5] Y. Hu et al., "SF2T: Self-supervised Fragment Finetuning of Video-LLMs for Fine-Grained Understanding," Comput. Vis. Found. J., pp. 29108–29117, 2025.

[6] N. N. Qonita, M. R. Handayani, and K. Umam, "Digital Forensic Chatbot Using DeepSeek LLM and NER for Automated Electronic Evidence Investigation," J. Tek. Inform., vol. 6, no. 3, pp. 1203–1216, 2025.

[7] R. Richo, "Sistem Identifikasi Informasi Expired Date Produk Kemasan Menggunakan Kolaborasi Metode Yolo-V11M Dan Paddleocr," J. Rekayasa Sist. Inf. dan Teknol., vol. 2, no. 3, pp. 886–900, 2025, doi: 10.70248/jrsit.v2i3.1719.

[8] S. Zhou et al., "EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering," CVPR Pap. J., pp. 3363–3373, 2025.

[9] R. Richo, "Analisis Keandalan YOLOv8m untuk Deteksi Varian Produk Kemasan Kotak pada Sistem Manajemen Kesediaan Stock," INFORMATICS Digit. Expert, vol. 2, pp. 124–131, 2024.

[10] Nur Wahyuningsih Ramadhani, E. Herianto, A. Fauzan, and M. Zubair, "PENGARUH MODEL PEMBELAJARAN KOOPERATIF TIPE GIVING QUESTION AND GETTING ANSWERS BERBASIS MEDIA AUDIO VISUAL TERHADAP HASIL BELAJAR PPKn PADA SISWA KELAS VIII DI SMPN 16 MATARAM," Pendas J. Ilm. Pendidik. Dasar, vol. 9, no. 02, pp. 3968–3977, 2024, doi: 10.2207/jjws.91.328.

[11] D. N. Pratomo, D. U. K. Putri, and A. Azhari, "Implementasi Optical Character Recognition berbasis Deep Learning untuk Ekstraksi Data Sertifikat Tanah," J. Inform. J. Pengemb. IT, vol. 7, no. 3, pp. 131–134, 2022.

[12] A. Yudertha and R. D. Putri, "Mapping Machine Learning Trends in Chemistry Research using LLM with Multi-Turn Prompting," Sistemasi, vol. 14, no. 2, p. 587, 2025, doi: 10.32520/stmsi.v14i2.4961.

[13] R. Richo, R. Y. Adhitya, M. K. Hasin, M. Syai'in, and E. Setiawan, "Eksplorasi Keandalan Sistem Sortir dan Klasifikasi Kecacatan Perekat Kemasan Menggunakan Arsitektur UNet-Inception Convolutional Neural Network," J. Elektron. dan Otomasi Ind., vol. 10, no. 3, pp. 321–333, 2023, doi: 10.33795/elkolind.v10i3.3835.

[14] F. Hibatulwafi and L. Laksmi, "Fenomena Penggunaan Generative AI dalam Perilaku Pencarian Informasi Praktisi Teknologi," Media Pustak., vol. 31, no. 2, pp. 141–155, 2024, doi: 10.37014/medpus.v31i2.5222.

[15] R. Richo, R. Yudha Adhitya, M. Khoirul Hasin, M. Syai'in, and E. Setiawan, "Analisis Pengaruh Optimizer pada Model CNN untuk Identifikasi Cacat pada Perekat Kemasan Optimizer," *J. Sisfotenika*, vol. 13, no. 2, pp. 217–229, 2023, [Online]. Available: http://sisfotenika.stmikpontianak.ac.id/index.php/ST