e-ISSN: 2460-870X | p-ISSN: 2085-1367

Implementasi dan Evaluasi Small Language Model untuk Chatbot CRM Menggunakan LoRA dan RAG

Implementation and Evaluation of Small Language Models for CRM Chatbots Using LoRA and RAG

Irawan Wingdes[1], Rendy Amy Saputra[2], Anne Putri Miranda[3]
[1,2,3] STMIK Pontianak, Jalan Merdeka No 372, Indonesia
irawan.wingdes@gmail.com

ABSTRAK

Adopsi large language models (LLM) dalam layanan customer relationship management (CRM) terus meningkat, tetapi penerapannya pada usaha mikro, kecil, dan menengah (UMKM) masih terkendala oleh keterbatasan sumber daya komputasi dan biaya. Penelitian ini mengkaji pemanfaatan small language model (SLM) sebagai alternatif efisien untuk chatbot layanan pelanggan. Pendekatan yang digunakan meliputi parameter-efficient fine-tuning (PEFT) dengan metode Low-Rank Adaptation (LoRA) serta perbandingan dengan retrieval-augmented generation (RAG). Penelitian menggunakan metode Design Science Research (DSR) dengan studi kasus pada usaha konveksi di Pontianak. Model Qwen2-1.5B-Instruct diadaptasi menggunakan LoRA dan diuji pada perangkat GPU 6GB. Evaluasi dilakukan secara kuantitatif dan kualitatif. Hasil menunjukkan bahwa model tanpa adaptasi memiliki performa rendah. Pendekatan LoRA menghasilkan performa paling stabil dengan intent accuracy hingga 95% - 100% dan hallucination rate 0%, sedangkan RAG meningkatkan pemahaman konteks namun kurang konsisten dalam output. Penelitian ini menyimpulkan bahwa fine-tuning efisien berbasis domain merupakan faktor penting dalam penerapan SLM untuk CRM UMKM.

Kata Kunci : Small Language Model, LoRA, CRM

ABSTRACT

The adoption of large language models (LLMs) in customer relationship management (CRM) has increased rapidly, but their use in micro, small, and medium enterprises (MSMEs) remains limited due to computational and cost constraints. This study explores the use of small language models (SLMs) as an efficient alternative for customer service chatbots, using parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA) and comparing it with retrieval-augmented generation (RAG). This research follows the Design Science Research (DSR) approach with a case study on a local garment production business in Pontianak. The Qwen2-1.5B-Instruct model is adapted using LoRA and deployed on a 6GB GPU. Evaluation is conducted through quantitative and qualitative methods. Results show that the base model performs poorly without adaptation. The LoRA approach achieves the most stable performance, with intent accuracy up to 95%–100% and 0% hallucination rate, while RAG improves contextual understanding but lacks output consistency. The study concludes that domain-specific efficient fine-tuning is crucial to enabling SLM-based CRM solutions for SMEs.

Keywords : Small Language Model, LoRA, CRM

Info Artikel :
Disubmit: *Direview:* *Diterima :*

1. PENDAHULUAN

Adopsi kecerdasan buatan khususnya large language models(LLM) dalam operasi bisnis meningkat pesat dan terus dioptimasi di Indonesia [1]. Adopsi model bahasa tersebut tidak lepas dari kelebihan arsitektur transformer [2] yang memungkinkan otomasi berbagai proses berbasis teks, termasuk analisis dokumen, pengolahan informasi pelanggan, serta sistem percakapan otomatis. Dalam konteks manajemen hubungan pelanggan (CRM), teknologi ini memungkinkan pengembangan layanan berbasis percakapan otomatis atau chatbot yang mampu memahami bahasa natural dan merespons pertanyaan

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

pengguna secara lebih fleksibel dibandingkan sistem tradisional berbasis aturan[3]. Implementasi teknologi tersebut pada akhirnya dapat digunakan secara luas meningkatkan efisiensi operasi.

Penerapan teknologi LLM dalam lingkungan usaha mikro, kecil hingga menengah (UMKM) masih menghadapi beberapa kendala dimana sebagian besar sistem LLM memerlukan infrastruktur komputasi maupun biaya yang tinggi. Selain membangun sendiri sistem komputasi tersebut, alternatif adalah pada layanan cloud yang juga diperhitungkan bulanan dari pemakaian yang pada akhirnya juga memerlukan biaya yang tidak sedikit hingga kendala privasi. Kelemahan komputasi pribadi hingga ketergantungan pada layanan eksternal tersebut akan bertolak belakang dengan ketersediaan sumber daya UMKM yang terbatas, khususnya di negara berkembang [4] Oleh karena itu, diperlukan penelitian mengenai penerapan LLM hemat yang sesuai dengan keterbatasan sumber daya komputasi atau small language models (SLM).

Dalam praktik, UMKM sering kali tidak menggunakan sistem CRM formal seperti perusahaan besar. Interaksi pelanggan lebih banyak berlangsung melalui media sosial dan aplikasi pesan instan seperti WhatsApp maupun Instagram direct message, yang dikenal sebagai Social CRM [5]. Pendekatan ini memungkinkan komunikasi cepat dan fleksibel, akan tetapi meningkatkan beban layanan bagi pemilik usaha atau staf layanan pelanggan. Sebagian besar pertanyaan pelanggan bersifat repetitif, mendorong kebutuhan sistem otomatisasi percakapan yang mampu menggantikan sebagian aktivitas layanan pelanggan. Chatbot berbasis aturan, meskipun telah lama digunakan, memiliki keterbatasan dalam menangani variasi bahasa informal dan pertanyaan di luar pola yang ditentukan.

Perkembangan SLM [6] menawarkan pendekatan baru dengan kemampuan memahami struktur semantik bahasa dan menghasilkan respons yang lebih fleksibel. Namun, model generik yang tersedia secara publik belum mampu memahami konteks bisnis spesifik, seperti katalog produk, maupun pengetahuan teknikal produk ataupun sapaan khusus dari pelanggan ke organisasi. Oleh karena itu, fine-tuning diperlukan agar model dapat beradaptasi dengan kebutuhan domain tertentu. Fine-tuning konvensional terhadap model besar memerlukan pembaruan miliaran parameter, yang tidak praktis untuk lingkungan dengan sumber daya terbatas. Untuk itu, parameter efficient fine tuning (PEFT) [7] dapat digunakan, dimana memungkinkan adaptasi model dengan biaya komputasi rendah.

Salah satu metode PEFT yang populer adalah Low-Rank Adaptation (LoRA), yang melatih sejumlah kecil parameter tambahan tanpa mengubah seluruh bobot model utama [8]. Metode ini banyak diadopsi karena mampu menurunkan kebutuhan komputasi dan memori secara signifikan, sehingga pelatihan dapat dilakukan pada perangkat keras terbatas. Selain efisien, LoRA memungkinkan model beradaptasi terhadap gaya bahasa dan format respons tertentu, menjadikannya ideal untuk skenario chatbot layanan pelanggan yang memerlukan konsistensi komunikasi [8]. Beberapa studi bahkan menunjukkan bahwa LoRA dapat mempercepat pelatihan sekaligus mempertahankan kualitas model pada tugas spesifik. Namun, karena hanya melatih parameter tambahan, efektivitas LoRA sangat bergantung pada kapasitas model dasar. Pada model berukuran kecil, LoRA cenderung hanya lebih efektif untuk penyesuaian format, gaya bahasa, dan klasifikasi terbatas, dibanding tugas generatif kompleks yang membutuhkan penalaran lebih luas [9].

Sebagai alternatif atau pelengkap fine-tuning, pendekatan Retrieval-Augmented Generation (RAG) juga seringkali diimplementasi untuk menguji kemampuan model dalam menjawab pertanyaan berbasis pengetahuan eksternal melalui penggabungan pencarian dokumen dan generasi teks [10]. Dalam sistem chatbot, RAG berpotensi meningkatkan relevansi jawaban terhadap informasi operasional yang dinamis, seperti katalog produk atau kebijakan layanan. Namun, efektivitas RAG pada SLM masih menghadapi tantangan besar karena performanya sangat bergantung pada kualitas retrieval dan kemampuan model dalam memanfaatkan konteks tambahan secara konsisten [11]. Keterbatasan kapasitas konteks dan daya penalaran pada SLM membuatnya rentan menghasilkan respons yang tidak konsisten atau halusinasi ketika proses retrieval tidak optimal atau konteks terlalu kompleks. Dalam sistem layanan pelanggan,

e-ISSN: 2460-870X | p-ISSN: 2085-1367

respons generatif yang tidak akurat ini menimbulkan risiko operasional yang tinggi karena model dapat memberikan informasi atau janji layanan yang tidak sesuai dengan kondisi nyata perusahaan.

Penelitian sebelumnya banyak membahas tentang PEFT berbasis LoRA dan RAG, tetapi mayoritas masih berfokus pada pengujian efisiensi di atas kertas, seperti pengukuran latensi, ukuran memori dan skor akurasi pada dataset laboratorium [9][12][13][14]. Namun, untuk pengujian di praktik nyata dengan komputasi lokal dan sumber daya terbatas, implementasi SLM masih tidak akurat dan rentan mengalami halusinasi akibat keterbatasan kapasitas konteks serta kegagalan pencarian dokumen pendukung [9][11]. Pada konteks penerapan CRM, adanya halusinasi ini akan secara langsung dapat memengaruhi service level karena SLM dapat memberikan janji atau informasi produk salah yang tidak dapat ditepati oleh organisasi. Pada akhirnya, hal ini dapat menimbulkan masalah baru dalam hubungan dengan pelanggan.

Selain masalah akurasi, definisi mengenai keterbatasan komputasi dalam literatur akademik seringkali tidak sesuai dengan realitas di lapangan. Di dalam berbagai eksperimen, komputasi minimal kerap kali didefinisikan sebagai penggunaan akselerator kelas data center tingkat rendah seperti NVIDIA Tesla P100 [15] atau NVIDIA T4 [16]. Secara praktik, perangkat tersebut memiliki harga yang tinggi (di atas 20 juta rupiah untuk kondisi bekas pakai), sehingga mungkin tidak terjangkau untuk skala UMKM, terutama di daerah berkembang seperti Pontianak. Berdasarkan pengamatan penulis, sebagian besar UMKM lokal masih beroperasi menggunakan perangkat tanpa GPU dedikasi. Oleh karena itu, asumsi penggunaan GPU kelas industri menjadi tidak realistis, baik dari sisi ketersediaan maupun anggaran biaya operasional.

Berdasarkan kesenjangan / gap tersebut, penelitian ini befokus pada kemampuan implementasi dalam kondisi UMKM nyata di Pontianak. Keterbatasan penelitian terdahulu akan diisi melalui dua aspek utama. Pertama, penelitian ini mengevaluasi implementasi model bahasa skala kecil langsung dalam konteks operasional nyata CRM UMKM. Kedua, proses pelatihan dilakukan menggunakan perangkat komputasi yang cukup mewakili kondisi aktual pelaku usaha kecil yang terbatas, yaitu GPU kelas konsumen NVIDIA GTX 1660 Super 6GB yang dirilis pada tahun 2019. Melalui pendekatan ini, sistem mengimplementasikan SLM dengan metode LoRA dan/atau RAG untuk mengevaluasi secara riil pilihan antara performa dan efisiensi di dalam keterbatasan sumber daya.

Secara spesifik, penelitian ini bertujuan merancang dan mengevaluasi chatbot CRM berbasis SLM yang diadaptasi menggunakan metode LoRA guna menilai efektivitasnya dalam mendukung layanan pelanggan pada usaha kecil. Studi kasus dilakukan pada sebuah usaha konveksi pakaian lokal di Pontianak yang melayani pelanggan ritel secara langsung dan menangani pertanyaan standar terkait produk serta pemesanan. Artefak sistem chatbot yang dihasilkan kemudian dievaluasi kemampuannya dalam mendukung interaksi pelanggan.

Melalui perancangan tersebut, penelitian ini memberikan tiga kontribusi utama: evaluasi integrasi SLM berbasis Qwen2-1.5B-Instruct [17], LoRA, dan RAG dalam konteks layanan CRM UMKM. Analisis komparatif antara pendekatan fine-tuning efisien (LoRA) dan retrieval based (RAG) pada model berkapasitas kecil untuk mengidentifikasi trade off performa dan efisiensi. Kelayakan implementasi model pada perangkat keras terbatas (GPU 6GB) dengan waktu pelatihan yang dibatasi kurang dari 12 jam, sesuai dengan batas operasional yang ditetapkan oleh objek penelitian. Dataset yang digunakan berasal dari data interaksi nyata objek penelitian yang berfokus pada domain spesifik usaha, sehingga seluruh proses pelatihan dan evaluasi dirancang untuk mencerminkan kondisi keterbatasan data dan komputasi yang umum dihadapi pelaku usaha kecil. Dengan demikian, pendekatan yang digunakan tidak hanya menilai performa model, tetapi juga menguji efektivitasnya dalam skenario implementasi nyata dalam sumber daya yang terbatas.

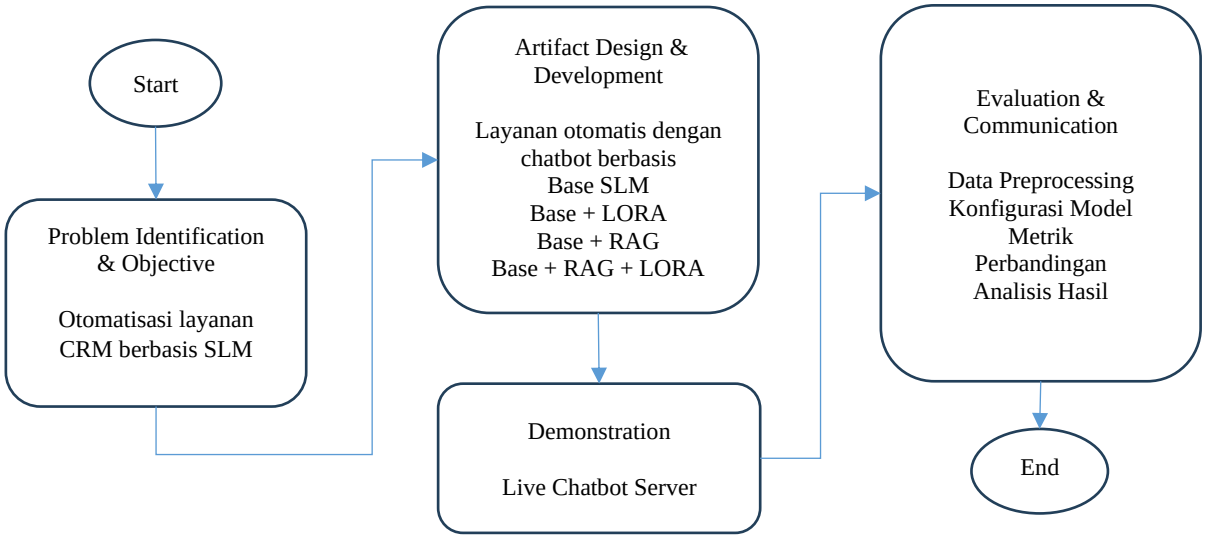
DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

2. METODE

Penelitian ini menggunakan pendekatan Design Science Research (DSR) [18] untuk merancang dan mengevaluasi artefak teknologi berupa sistem chatbot berbasis small language model untuk aktivitas CRM pada usaha mikro. Objek studi kasus usaha adalah konveksi kaos lokal di Pontianak yang melayani jual beli di daerah Pontianak dengan umur operasi di atas 15 tahun.

Identifikasi masalah dilakukan melalui studi kualitatif [19] pada usaha. Tujuan solusi dirumuskan untuk mengembangkan sistem chatbot yang mampu menjawab pertanyaan repetitif pelanggan secara otomatis, menyadur niat pelanggan untuk dilanjutkan dengan langkah standar, menggunakan model bahasa berskala kecil yang dapat dijalankan pada perangkat keras dengan graphical processing unit berkapasitas 6 giga byte yang dimiliki pelaku usaha. Perancangan sistem chatbot dilakukan dengan mengadaptasi small language model menggunakan teknik PEFT dan LORA.

Model dasar yang digunakan adalah Qwen Instruct 1.5B, dengan kapasitas sekitar 1,5 miliar parameter dengan pertimbangan dapat dijalankan sesuai perangkat keras usaha. Adaptasi model dilakukan dengan melatih sejumlah kecil parameter tambahan tanpa memperbarui seluruh bobot model utama, sehingga proses fine-tuning menjadi lebih efisien secara komputasi. Selain itu, sistem juga akan mengintegrasikan pendekatan RAG menjawab pertanyaan berbasis informasi produk (bersifat dinamis) yang tersimpan dalam basis pengetahuan usaha [7][8][10].



Gambar 1. Langkah Penelitian
(Sumber: Data Olahan)

Implementasi sistem chatbot dilakukan dalam bentuk percakapan langsung dengan pelanggan yang mensimulasikan interaksi layanan pelanggan pada usaha konveksi pakaian. Sistem dioperasikan menggunakan Telegram bot yang berjalan pada server lokal (perangkat komputasi pemilik usaha).

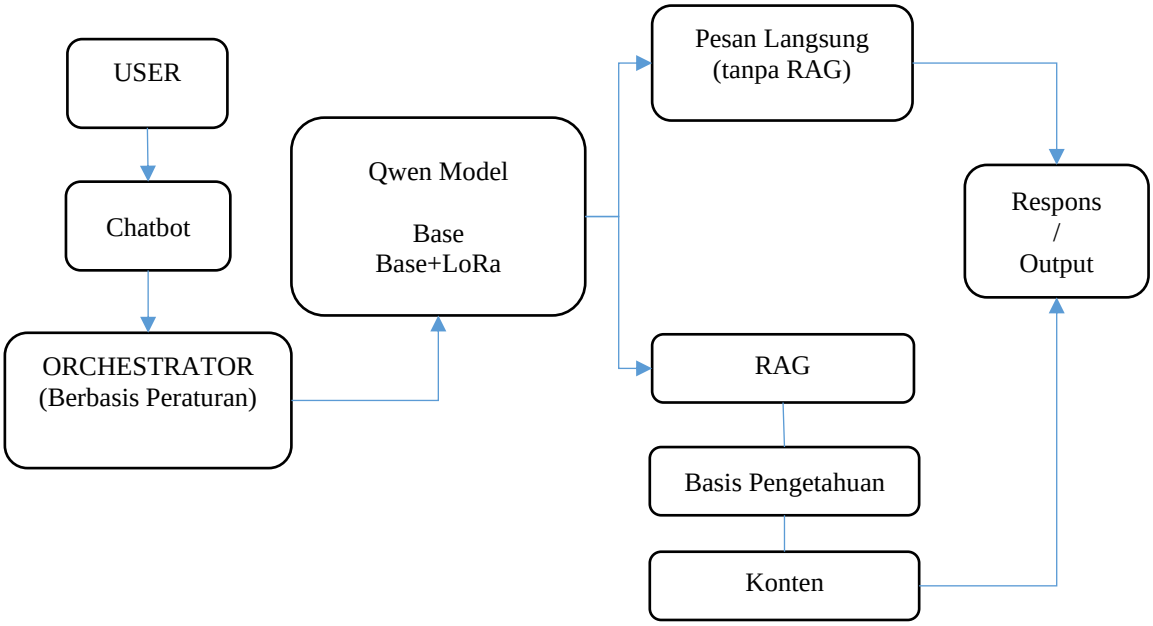
Dataset percakapan pelanggan disusun dari riwayat percakapan WhatsApp toko, ringkasan pertanyaan repetitif konsumen oleh staf layanan pelanggan, serta variasi pertanyaan yang diproduksi dengan algoritma secara otomatis untuk meningkatkan keragaman bahasa. Setiap data percakapan diberi label intent yang merepresentasikan tujuan utama pertanyaan pelanggan. Terdapat 6 intent yang dilatih yaitu product_info, shipping_info, payment_info, order_status, human_support, dan greeting.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Selain proses pelatihan model, penelitian ini juga merancang arsitektur sistem chatbot secara menyeluruh untuk memastikan integrasi antara komponen pemahaman bahasa, pengelolaan logika bisnis, serta antarmuka pengguna. Arsitektur sistem terdiri dari tiga komponen utama, yaitu (1) modul pemrosesan bahasa alami berbasis small language model, (2) modul orkestrasi berbasis aturan, dan (3) modul integrasi antarmuka percakapan.

Modul pemrosesan bahasa alami bertugas untuk menerima input teks dari pengguna dan mengklasifikasikan intent menggunakan model yang telah diadaptasi dengan metode LoRA. Output dari modul ini berupa struktur JSON sederhana yang berisi label intent. Desain ini bertujuan untuk menjaga kompleksitas model tetap rendah, sehingga seluruh proses inferensi dapat dijalankan secara efisien pada perangkat keras dengan keterbatasan sumber daya.

Selanjutnya, modul orkestrasi berbasis aturan bertanggung jawab untuk menentukan respons sistem berdasarkan intent yang dihasilkan. Setiap intent dipetakan ke skenario respons tertentu yang telah didefinisikan berdasarkan standar operasional usaha. Pada tahap ini, sistem tidak hanya menghasilkan respons statis, tetapi juga dapat memicu alur lanjutan seperti permintaan informasi tambahan atau pengalihan ke staf manusia. Pendekatan ini memungkinkan pemisahan yang jelas antara komponen pembelajaran mesin dan logika bisnis, sehingga memudahkan proses pemeliharaan dan pengembangan sistem.



Gambar 2. Arsitektur Chatbot
(Sumber: Data Olahan)

Pada konfigurasi yang menggunakan pendekatan RAG, modul tambahan berupa komponen retrieval diintegrasikan sebelum proses generasi respons. Modul ini bertugas untuk mengambil dokumen yang relevan dari basis pengetahuan menggunakan teknik pencarian berbasis embedding. Dokumen yang terpilih kemudian disisipkan ke dalam prompt sebagai konteks tambahan sebelum dikirim ke model bahasa. Integrasi ini dirancang secara modular sehingga dapat diaktifkan atau dinonaktifkan tanpa mengubah struktur utama sistem.

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

Modul integrasi antarmuka percakapan diimplementasikan menggunakan Telegram Bot API sebagai prototipe sistem (gambar 2). Pemilihan platform ini didasarkan pada kemudahan integrasi serta fleksibilitas dalam pengujian. Sistem dirancang untuk mendukung transisi ke platform lain seperti WhatsApp Business API dengan perubahan minimal.

Evaluasi performa sistem dilakukan secara kuantitatif dan kualitatif. Evaluasi kuantitatif dilakukan dengan 2 set pertanyaan yaitu formal dan informal. metrik evaluasi adalah intent accuracy, exact match, F1 score, dan hallucination rate untuk mengukur kemampuan chatbot dalam memahami dan merespons pertanyaan pelanggan secara tepat dan akurat. Intent accuracy digunakan untuk mengukur kemampuan sistem dalam mengidentifikasi kategori intent dari pertanyaan pengguna [10]. Exact match dan F1 score digunakan untuk mengevaluasi kesesuaian antara jawaban yang dihasilkan sistem dengan jawaban referensi, sebagaimana umum digunakan dalam evaluasi sistem question answering [20][21]. Selain itu, hallucination rate digunakan untuk mengukur proporsi respons yang mengandung informasi yang tidak didukung oleh basis pengetahuan sistem.

Evaluasi kualitatif dilakukan melalui interaksi langsung dengan partisipan yang terdiri dari pemilik usaha, 2 staf layanan pelanggan, dan 5 konsumen. Partisipan berinteraksi dengan chatbot menggunakan skenario percakapan yang mencakup pertanyaan formal dan informal, kemudian memberikan umpan balik terkait kejelasan jawaban, kesesuaian informasi, kemudahan penggunaan, serta potensi pemanfaatan chatbot dalam operasional layanan pelanggan. Hasil evaluasi dianalisis untuk mengidentifikasi kelebihan dan keterbatasan sistem dalam konteks penggunaan nyata pada usaha mikro. Penelitian ini dilaksanakan selama empat bulan pada akhir tahun 2025 di lokasi operasi tempat usaha di Tanjungpura.

3. HASIL DAN PEMBAHASAN

Penelitian dimulai dengan identifikasi masalah melalui studi kualitatif terhadap percakapan antara pelanggan dan pihak usaha. Proses ini menghasilkan dataset percakapan pelanggan yang kemudian dianalisis untuk mengidentifikasi pola pertanyaan yang dominan. Dataset tersebut selanjutnya ditransformasikan menjadi data pelatihan dalam format sederhana dengan mempertimbangkan keterbatasan kapasitas model yang digunakan, yaitu sekitar 1,5 miliar parameter.

Model dirancang untuk berperan secara spesifik dalam melakukan klasifikasi intent dari pertanyaan pelanggan. Setelah intent teridentifikasi, proses respons ditangani oleh sistem orkestrasi berbasis aturan (rule based orchestrator) yang mengacu pada standar operasional perusahaan. Pendekatan ini bertujuan untuk memisahkan kemampuan pemahaman bahasa alami dari kontrol logika bisnis, sehingga meningkatkan konsistensi dan reliabilitas sistem [21].

Pendekatan tersebut juga memberikan efisiensi komputasi sekaligus menjaga kesesuaian respons dengan kebijakan perusahaan. Tanpa proses adaptasi, model dasar cenderung menghasilkan respons generik terhadap pertanyaan seperti “Ada kaos ukuran XL?”. Setelah dilakukan fine-tuning, model mampu menghasilkan respons yang lebih terpersonalisasi (“Terima kasih atas minatnya kak, ukuran XL ada ya kak”) dan sesuai dengan gaya komunikasi perusahaan, sehingga meningkatkan kesesuaian terhadap kebutuhan informasi pengguna.

Berdasarkan hasil analisis dataset, diperoleh enam kategori intent utama, yaitu product_info, shipping_info, payment_info, order_status, human_support, dan greeting. Klasifikasi ini merepresentasikan distribusi kebutuhan informasi pelanggan dalam konteks operasional usaha.

Pengembangan model dilakukan menggunakan bahasa pemrograman Python dengan model dasar Qwen2-1.5B-Instruct. Model dijalankan pada perangkat keras kelas konsumen berbasis AMD AM4 dengan GPU NVIDIA GeForce GTX 1660 Super (6 GB VRAM). Untuk memungkinkan proses fine-tuning pada sumber daya terbatas, model dimuat menggunakan kuantisasi 8-bit melalui pustaka

e-ISSN: 2460-870X | p-ISSN: 2085-1367

BitsAndBytes. Pendekatan ini terbukti efektif dalam mengurangi kebutuhan memori tanpa penurunan performa yang signifikan [22], sehingga waktu pelatihan berkurang dari sekitar 4 jam menjadi 1,5–2 jam.

Proses fine-tuning dilakukan menggunakan metode Low-Rank Adaptation (LoRA) dengan konfigurasi rank 8, alpha 16, dan dropout 0,05. Adaptasi diterapkan pada modul proyeksi attention (q_proj, k_proj, v_proj, o_proj). Jumlah parameter yang dilatih sebesar 2.179.072 dari total 1.545.893.376 parameter (sekitar 0,141%). Pendekatan ini memungkinkan efisiensi pelatihan tanpa mengorbankan performa secara signifikan [8]. Dataset dibagi menjadi data pelatihan (2.989), validasi (640), dan pengujian (641). Contoh dataset pelatihan pada tabel 1.

Tabel 1. Dataset Sampel Pelatihan

Instruksi	**Input**	**Output**	**Intent**
Identifikasi Intent dari user dan kembalikan output dalam format JSON.	cek kaos L	{"intent": "product_info"}	product_info
	kemeja biru min?	{"intent": "product_info"}	product_info
	saya ingin kemeja custom	{"intent": "human_support"}	human_support
	desain sendiri kah	{"intent": "human_support"}	human_support
	kurir toko bisa?	{"intent": "shipping_info"}	shipping_info
	pake Grab gak?	{"intent": "shipping_info"}	shipping_info
	pesananku sampai mana ya?	{"intent": "order_status"}	order_status
	halo mau tanya pesanan ku	{"intent": "order_status"}	order_status
	bisa bayar pakai tf?	{"intent": "payment_info"}	payment_info
	bayar via transfer bisa ga sih?	{"intent": "payment_info"}	payment_info
	Pagi	{"intent": "greeting"}	greeting
	Siang	{"intent": "greeting"}	greeting

Untuk mengevaluasi kemampuan model dalam memanfaatkan pengetahuan eksternal, digunakan pendekatan Retrieval-Augmented Generation (RAG), yang mengintegrasikan proses retrieval berbasis embedding dengan generasi respons berbasis model bahasa [10]. Basis pengetahuan terdiri dari dokumen operasional usaha yang mencakup informasi produk, pengiriman, pembayaran, dan layanan pelanggan.

Dokumen dipecah menjadi unit informasi dan diubah menjadi representasi vektor menggunakan model paraphrase-multilingual-MiniLM-L12-v2. Pemilihan model embedding ini didasarkan pada efisiensi komputasi dan kemampuannya dalam merepresentasikan kesamaan semantik multibahasa secara akurat [23]. Kesamaan antara pertanyaan pengguna dan dokumen dihitung menggunakan cosine similarity, yang umum digunakan dalam pencarian semantik karena kestabilannya dalam mengukur kedekatan arah vektor [24].

Dua dokumen dengan skor kesamaan tertinggi dipilih sebagai konteks tambahan. Pemilihan top-k = 2 didasarkan pada pertimbangan pilihan antara relevansi informasi dan keterbatasan kapasitas konteks model. Penambahan konteks yang berlebihan dapat meningkatkan kompleksitas prompt dan menurunkan

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

efektivitas pemrosesan pada model berukuran kecil [25]. Oleh karena itu, strategi ini bertujuan untuk mempertahankan informasi yang relevan tanpa membebani kapasitas pemahaman model.

Evaluasi sistem dilakukan menggunakan dua kategori pertanyaan, yaitu formal dan informal. Pertanyaan formal menggunakan bahasa baku sesuai EYD, sedangkan pertanyaan informal mencerminkan variasi bahasa percakapan sehari-hari. Setiap kategori terdiri dari 30 pertanyaan untuk masing-masing intent, sehingga total 180 pertanyaan per model. Sampel dari pengujian pada tabel 2.

Tabel 2. Dataset Sampel Pengujian

Pertanyaan	**Intent**	**Kategori**
Apakah tersedia kaos warna hitam?	product_info	test_formal
Apakah kemeja Oxford tersedia ukuran S?	product_info	test_formal
kaos warna pink size L ada gak	product_info	test_informal
tshirt L ready gak min	product_info	test_informal
Apakah tersedia pengiriman Maxim?	shipping_info	test_formal
Apakah bisa pengiriman menggunakan Gojek?	shipping_info	test_formal
kirim via Gojek ya?	shipping_info	test_informal
gojek bisa?	shipping_info	test_informal
bisa kirim pake Maxim gak	shipping_info	test_informal
Apakah bisa membayar menggunakan QRIS?	payment_info	test_formal
Apakah bisa membayar dengan transfer bank?	payment_info	test_formal
bisa bayar pake qris gak?	payment_info	test_informal
bisa bayar pake tf bank gak?	payment_info	test_informal
Halo	greeting	test_formal
Assalamualaikum	greeting	test_formal
haloooo	greeting	test_informal
permisiiisss	greeting	test_informal
Bagaimana status pesanan saya?	order_status	test_formal
Apakah pesanan saya sudah dikemas?	order_status	test_formal
cek status pesanan dong	order_status	test_informal
pesanan saya sudah dikemas belum ya?	order_status	test_informal
Saya ingin berbicara dengan admin	human_support	test_formal
Saya ingin melakukan komplain	human_support	test_formal
bantu dong min	human_support	test_informal
mau komplain nih	human_support	test_informal

Pada tahap demonstrasi, sistem chatbot dioperasikan pada server lokal yang berada di lingkungan operasional usaha serta terhubung ke jaringan internet. Dalam fase ini, chatbot diintegrasikan dengan platform Telegram sebagai prototipe antarmuka interaksi dengan pelanggan. Untuk tahap implementasi

e-ISSN: 2460-870X | p-ISSN: 2085-1367

penuh, sistem direncanakan untuk diintegrasikan dengan WhatsApp Business agar dapat digunakan secara langsung dalam aktivitas layanan pelanggan sehari-hari.

Untuk konfigurasi tanpa RAG, evaluasi dilakukan secara otomatis menggunakan metrik intent accuracy, exact match, F1-score, dan hallucination rate. Hallucination rate didefinisikan sebagai proporsi output yang tidak valid secara struktural atau tidak sesuai dengan label intent. Evaluasi dilakukan menggunakan pustaka scikit-learn [26].

Pada konfigurasi sistem yang menggunakan pendekatan RAG, evaluasi tidak dilakukan secara otomatis, melainkan melalui penilaian manual. Hal ini disebabkan oleh karakteristik output model yang bersifat generatif, sehingga tidak selalu dapat dievaluasi secara langsung menggunakan metrik berbasis kesesuaian label tanpa mempertimbangkan konteks. Oleh karena itu, setiap respons dianalisis secara kualitatif oleh evaluator manusia berdasarkan tiga kriteria utama, yaitu kesesuaian dengan intent, kualitas informasi yang disampaikan, serta indikasi adanya informasi yang bersifat halusinatif. Hasil evaluasi terhadap empat konfigurasi model pada skenario pertanyaan formal disajikan pada Tabel 3.

Tabel 3. Hasil Pengujian 4 Model Pertanyaan Formal

Model	Intent Accuracy	Exact Match	F1 Score	Hallucination
Qwen	10.56	10.56	0.052	71.11
qwen rag	98.3	-	-	26.6
qwen lora rag	100	-	-	33.9
qwen lora	95	95	0.948	0

Hasil pengujian pada Tabel 3 menunjukkan perbedaan performa yang kontras antar konfigurasi model. Model dasar tanpa proses adaptasi (Qwen) hanya mencapai intent accuracy sebesar 10,56% dengan hallucination rate sebesar 71,11%. Nilai ini mengindikasikan bahwa sebagian besar output yang dihasilkan tidak sesuai dengan label intent yang diharapkan, serta banyak respons yang gagal disadur ke dalam format JSON yang valid. Hal ini menunjukkan keterbatasan model bahasa generik dalam menangani tugas ekstraksi intent yang bersifat terstruktur dalam konteks spesifik usaha.

Penerapan pendekatan RAG pada model dasar menghasilkan peningkatan intent accuracy yang sangat signifikan menjadi 98,3%. Peningkatan ini menunjukkan bahwa penambahan konteks eksternal melalui mekanisme retrieval berkontribusi dalam membantu model mengaitkan pertanyaan pengguna dengan informasi yang relevan. Namun, hallucination rate masih tercatat sebesar 26,6%, yang menunjukkan bahwa meskipun pemahaman intent meningkat, konsistensi format output belum sepenuhnya terjaga. Model masih menghasilkan respons yang tidak selalu sesuai dengan struktur JSON yang diharapkan.

Model yang diadaptasi menggunakan LoRA menunjukkan karakteristik performa yang berbeda. Konfigurasi Qwen + LoRA mencapai intent accuracy sebesar 95% dan exact match sebesar 95%, dengan F1-score sebesar 0,948 serta hallucination rate sebesar 0%. Nilai exact match yang identik dengan intent accuracy menunjukkan bahwa sebagian besar prediksi tidak hanya benar secara label, tetapi juga sesuai secara struktural. Ketiadaan hallucination mengindikasikan bahwa seluruh output yang dihasilkan dapat dipindahkan dengan baik ke dalam format JSON yang valid.

Lebih lanjut, kombinasi Qwen + LoRA + RAG menghasilkan intent accuracy tertinggi sebesar 100%. Hal ini menunjukkan bahwa seluruh pertanyaan dalam skenario pengujian berhasil diklasifikasikan dengan benar. Namun, konfigurasi ini menghasilkan hallucination rate sebesar 33,9%, yang lebih tinggi dibandingkan dengan penggunaan RAG tanpa LoRA. Kondisi ini menunjukkan bahwa meskipun

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

integrasi LoRA dan RAG meningkatkan ketepatan klasifikasi intent, terdapat peningkatan proporsi output yang tidak konsisten secara struktural.

Secara keseluruhan, perbedaan antar konfigurasi pada Tabel 3 menunjukkan variasi yang jelas antara kemampuan klasifikasi intent dan stabilitas output, di mana setiap pendekatan memberikan karakteristik performa yang berbeda pada kedua aspek tersebut.

Pengujian untuk pertanyaan informal dapat dilihat pada tabel 4. Hasil evaluasi pada Tabel 4 menunjukkan bahwa performa model dasar Qwen2-1.5B-Instruct mengalami penurunan lebih lanjut pada skenario pertanyaan dengan bahasa informal. Model hanya mencapai intent accuracy sebesar 8,33%, exact match sebesar 8,33%, serta F1-score sebesar 0,026, dengan hallucination rate sebesar 41,67%. Nilai exact match yang identik dengan intent accuracy menunjukkan bahwa hanya sebagian kecil prediksi yang sesuai baik dari sisi label maupun struktur. Selain itu, nilai F1-score yang sangat rendah mengindikasikan ketidakseimbangan antara presisi dan recall dalam proses klasifikasi intent.

Tabel 4. Hasil Pengujian 4 Model Pertanyaan Informal

Model	Intent Accuracy	Exact Match	F1 Score	Hallucination
qwen	8.33	8.33	0.026	41.67
qwen rag	67.7	-	-	25
qwen lora rag	76.1	-	-	19.9
qwen lora	87.78	87.78	0.873	0

Penerapan RAG pada model dasar menghasilkan peningkatan performa yang cukup signifikan, dengan intent accuracy sebesar 67,7% dan hallucination rate sebesar 25%. Peningkatan ini menunjukkan bahwa mekanisme retrieval membantu model dalam mengaitkan pertanyaan pengguna dengan konteks informasi yang relevan.

Kombinasi LoRA dan RAG menghasilkan peningkatan performa lebih lanjut dengan intent accuracy sebesar 76,1% dan hallucination rate sebesar 19,9%. Dibandingkan dengan penggunaan RAG saja, konfigurasi ini menunjukkan penurunan hallucination rate serta peningkatan akurasi klasifikasi, yang mengindikasikan adanya perbaikan dalam pemahaman intent sekaligus peningkatan konsistensi output, meskipun masih terdapat proporsi respons yang tidak valid.

Model yang diadaptasi menggunakan LoRA tanpa RAG menunjukkan performa paling stabil. Konfigurasi ini mencapai intent accuracy sebesar 87,78%, exact match sebesar 87,78%, serta F1-score sebesar 0,873, dengan hallucination rate sebesar 0%. Kesetaraan antara intent accuracy dan exact match menunjukkan bahwa prediksi yang dihasilkan tidak hanya tepat secara label, tetapi juga konsisten dalam format JSON yang diharapkan. Selain itu, nilai F1-score yang tinggi menunjukkan keseimbangan yang baik antara presisi dan recall dalam klasifikasi intent.

Hasil pada Tabel 4 menunjukkan bahwa seluruh konfigurasi mengalami penurunan performa dibandingkan skenario formal, tetapi dengan tingkat penurunan yang berbeda. Variasi ini mencerminkan perbedaan kemampuan setiap pendekatan dalam menangani karakteristik bahasa yang tidak terstruktur, baik dari sisi pemahaman intent maupun konsistensi format output.

Dari sisi penggunaan sumber daya, pendekatan LoRA menunjukkan efisiensi yang signifikan dibandingkan fine-tuning konvensional. Dengan hanya melatih sekitar 0,141% dari total parameter model, kebutuhan memori selama proses pelatihan dan inferensi dapat ditekan dengan efektif. Hal ini

e-ISSN: 2460-870X | p-ISSN: 2085-1367

memungkinkan model dijalankan pada GPU dengan kapasitas 6GB tanpa memerlukan teknik distribusi atau offload tambahan. Selain itu, penggunaan kuantisasi 8-bit juga berkontribusi dalam mengurangi footprint memori tanpa menurunkan performa secara signifikan. Dari sisi waktu, pelatihan hanya membutuhkan waktu 1 hingga 2 jam, menurun dibandingkan 4 jam.

Dalam konteks operasional, sistem chatbot yang dikembangkan menunjukkan potensi untuk menggantikan sebagian besar pertanyaan repetitif pelanggan. Berdasarkan hasil observasi selama tahap demonstrasi, sekitar 60 hingga 70% pertanyaan yang masuk dapat ditangani secara otomatis oleh sistem tanpa intervensi manusia. Hal ini berpotensi secara langsung mengurangi beban kerja staf layanan pelanggan, terutama pada jam operasional sibuk.

Namun, terdapat beberapa keterbatasan yang teridentifikasi selama implementasi. Pada konfigurasi berbasis RAG, variasi struktur output yang dihasilkan model menyebabkan kesulitan dalam integrasi dengan sistem orkestrasi. Respons yang tidak konsisten dalam format JSON memerlukan penanganan tambahan, seperti validasi dan normalisasi output, yang pada akhirnya meningkatkan kompleksitas sistem secara keseluruhan.

Selain itu, pada beberapa kasus pertanyaan ambigu atau multi-intent, model masih mengalami kesulitan dalam menentukan intent utama secara konsisten. Kondisi ini terutama terjadi pada pertanyaan yang menggabungkan beberapa tujuan sekaligus, seperti pertanyaan terkait produk yang diikuti dengan permintaan informasi pengiriman. Hal ini menunjukkan bahwa pendekatan klasifikasi intent tunggal masih memiliki keterbatasan dalam merepresentasikan kompleksitas percakapan nyata.

Dari perspektif pengalaman pengguna, hasil evaluasi kualitatif menunjukkan bahwa konsistensi gaya bahasa menjadi faktor penting dalam penerimaan sistem. Model yang telah diadaptasi dengan LoRA dinilai lebih sesuai dengan karakter komunikasi usaha, sehingga memberikan kesan interaksi yang lebih natural dan personal. Sebaliknya, model tanpa fine-tuning cenderung menghasilkan respons yang terlalu formal atau generik, sehingga kurang sesuai dengan ekspektasi pelanggan.

Temuan ini mengindikasikan bahwa keberhasilan implementasi chatbot pada UMKM tidak hanya ditentukan oleh akurasi teknis, tetapi juga oleh kesesuaian dengan konteks sosial dan budaya komunikasi pelanggan. Oleh karena itu, adaptasi domain tidak hanya berperan dalam meningkatkan performa model, tetapi juga dalam membentuk pengalaman pengguna yang lebih baik.

Hasil evaluasi kuantitatif pada Tabel 3 dan Tabel 4 menunjukkan pola yang konsisten terkait perbedaan karakteristik performa antar pendekatan. Model dasar tanpa adaptasi menunjukkan performa yang sangat rendah pada kedua skenario, baik formal maupun informal. Hal ini menegaskan bahwa model bahasa generik tidak mampu menangani tugas ekstraksi intent dalam domain spesifik tanpa proses adaptasi tambahan.

Pendekatan Retrieval-Augmented Generation (RAG) menunjukkan peningkatan signifikan pada intent accuracy, terutama pada skenario formal. Namun, hallucination rate yang tetap tinggi pada kedua skenario menunjukkan bahwa peningkatan pemahaman tidak diikuti oleh konsistensi output. Secara operasional, kondisi ini mengindikasikan bahwa pendekatan RAG pada model berukuran kecil belum cukup andal untuk digunakan dalam sistem layanan pelanggan yang membutuhkan tingkat keberhasilan mutlak yang tinggi (misalnya service level 95% dalam menanggapi kemauan pelanggan).

Sebaliknya, pendekatan Low-Rank Adaptation (LoRA) menunjukkan performa yang paling stabil pada seluruh metrik evaluasi. Nilai intent accuracy, exact match, dan F1-score yang tinggi, serta hallucination rate sebesar 0%, menunjukkan bahwa model tidak hanya mampu memahami intent dengan baik, tetapi juga menghasilkan keluaran yang konsisten secara struktural. Temuan ini mengindikasikan bahwa fine-tuning berbasis domain memiliki peran dominan dalam menyesuaikan representasi internal

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

model terhadap pola bahasa spesifik, bahkan dengan jumlah parameter yang dilatih relatif kecil [8]. Temuan tersebut juga menunjukkan bahwa dalam kondisi keterbatasan data dan sumber daya komputasi, pendekatan fine-tuning berbasis LoRA lebih sesuai untuk implementasi praktis dibandingkan pendekatan berbasis retrieval.

Gambar 3. Pengujian Chatbot Base + LoRA
(Sumber: Data Olahan)

Perbandingan antara skenario formal dan informal menunjukkan bahwa seluruh konfigurasi mengalami penurunan performa pada bahasa tidak terstruktur. Namun, penurunan tersebut paling kecil terjadi pada model dengan LoRA, yang menunjukkan kemampuan adaptasi terhadap variasi bahasa percakapan (gambar 3). Hal ini sejalan dengan temuan bahwa fine-tuning berbasis domain meningkatkan kehandalan terhadap distribusi bahasa yang berbeda dari data pelatihan awal [27].

Temuan kualitatif pada tahap demonstrasi memperkuat hasil kuantitatif tersebut. Dalam skenario interaksi nyata, hanya model dengan LoRA yang mampu memberikan respons secara konsisten sesuai dengan standar layanan pelanggan, baik dari sisi relevansi maupun struktur. Model berbasis RAG, meskipun mampu menghasilkan respons yang informatif, menunjukkan variasi struktur dan konsistensi format yang berpotensi mengganggu integrasi dengan sistem orkestrasi. Kondisi ini mengindikasikan bahwa pada model berukuran sekitar 1,5 miliar parameter, pendekatan RAG belum cukup andal untuk digunakan secara langsung dalam sistem layanan pelanggan yang membutuhkan konsistensi tinggi.

Lebih lanjut, integrasi LoRA dan RAG tidak selalu menghasilkan peningkatan performa yang linear. Meskipun kombinasi ini mencapai intent accuracy tertinggi pada skenario formal, peningkatan hallucination rate menunjukkan perlunya pilihan antara akurasi klasifikasi dan stabilitas output. Fenomena ini mengindikasikan bahwa penambahan konteks eksternal dapat meningkatkan kompleksitas yang harus diproses oleh model, sementara kapasitas model kecil memiliki keterbatasan dalam

e-ISSN: 2460-870X | p-ISSN: 2085-1367

mengelola konteks tersebut secara optimal [25]. Hal ini menunjukkan bahwa pada model berukuran kecil, penambahan kompleksitas melalui integrasi RAG justru dapat menurunkan reliabilitas sistem, sehingga kurang sesuai untuk implementasi pada lingkungan dengan kehandalan dan stabilitas tinggi seperti layanan pelanggan UMKM.

Secara keseluruhan, temuan penelitian ini menunjukkan bahwa implementasi small language model (SLM) berukuran 1.5B parameter dalam layanan pelanggan UMKM paling layak digunakan untuk tugas terbatas, khususnya klasifikasi intent berbasis struktur. Pendekatan fine-tuning melalui LoRA terbukti menjadi faktor kunci dalam mencapai akurasi, stabilitas, dan konsistensi output. Sebaliknya, pendekatan generatif seperti RAG belum menunjukkan reliabilitas yang memadai untuk menghasilkan respons operasional secara penuh. Oleh karena itu, implementasi praktis SLM saat ini harus memprioritaskan keandalan dibanding kecanggihan, dengan memfokuskan peran model sebagai komponen pemahaman intent, bukan sebagai generator respons otomatis.

4. KESIMPULAN

a. Penerapan small language model (SLM) dalam konteks UMKM untuk layanan CRM dapat diimplementasikan dengan perangkat keras kelas konsumen yang memiliki keterbatasan sumber daya komputasi.
b. Model bahasa generik tanpa adaptasi menunjukkan performa yang sangat rendah dalam tugas klasifikasi intent, sehingga tidak memadai untuk digunakan langsung.
c. Pendekatan parameter-efficient fine-tuning (PEFT) dengan metode LoRA paling efektif dalam meningkatkan performa model, dengan capaian intent accuracy tinggi, F1-score tinggi, serta konsistensi output (hallucination rate 0%), baik pada pengujian skenario bahasa formal maupun informal.
d. Pendekatan retrieval-augmented generation (RAG) mampu meningkatkan kemampuan model dalam memahami konteks pertanyaan melalui penambahan informasi eksternal, tetapi masih menunjukkan keterbatasan dalam akurasi dan respon output, sehingga belum memadai untuk digunakan secara langsung dalam implementasi layanan pelanggan yang memiliki service level tinggi.
e. Kombinasi LoRA dan RAG menghasilkan peningkatan akurasi klasifikasi pada beberapa skenario, tetapi meningkatkan hallucination rate, sehingga kurang memadai untuk implementasi nyata.
f. Dalam implementasi, pendekatan berbasis SLM + LoRA saja menunjukkan performa paling stabil dan paling sesuai untuk kebutuhan layanan CRM objek penelitian.
g. Evaluasi kualitatif menunjukkan bahwa chatbot berbasis SLM dengan adaptasi domain mampu memberikan respons yang relevan dan sesuai dengan gaya komunikasi usaha tetapi dengan kemampuan yang terbatas, sehingga memerlukan kompromi atau prioritas tujuan apabila diimplementasikan.
h. Secara keseluruhan, penelitian ini menunjukkan bahwa adaptasi domain melalui fine-tuning LoRA merupakan faktor penting dalam keberhasilan penerapan SLM pada layanan CRM UMKM, sementara RAG lebih cocok digunakan sebagai mekanisme pelengkap yang masih memerlukan pengembangan lebih lanjut untuk mencapai tingkat layanan yang diperlukan.
i. Pada penelitian selanjutnya, integrasi pendekatan RAG dapat dikembangkan pada model dengan parameter lebih besar, untuk mengurangi hallucination rate maupun kemampuan pemahaman dan retrieval yang lebih tinggi.
j. Spesifikasi komputer yang digunakan dapat ditingkatkan agar dapat memuat model yang lebih besar dan kecepatan pelatihan yang lebih tinggi. Selain itu, dataset pelatihan pada akhirnya dapat diperluas baik dari sisi jumlah maupun keragaman bahasa, termasuk variasi dialek, slang, dan kesalahan penulisan, untuk meningkatkan robustness model terhadap kondisi percakapan nyata.

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.351-365

UCAPAN TERIMA KASIH

Unit Konveksi Tanjungpura dan STMIK Pontianak atas dukungan penelitian .

REFERENSI

[1] Kurnia, E. (2025). AI makin terjangkau, investasi di Indonesia bisa lebih optimal. Diambil dari https://www.kompas.id/artikel/ai-makin-terjangkau-investasi-di-indonesia-bisa-lebih-optimal

[2] Mohiuddin, K., et al. (2023). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems*, hal. 1–15.

[3] Guizani, S., Mazhar, T., Shahzad, T., Ahmad, W., Bibi, A., & Hamam, H. (2025). A systematic literature review to implement large language model in higher education: Issues and solutions. *Discover Education*, 4(1).

[4] Moenks, N., Penava, P., & Buettner, R. (2025). A systematic literature review of large language model applications in industry. *IEEE Access*, 13, hal. 160010–160033.

[5] Mustafa, H., et al. (2023). The impact of using WhatsApp business (API) in marketing for small business. *Proceedings of the 2023 10th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, hal. 1–9.

[6] Belcak, P., et al. (2025). Small language models are the future of agentic AI. *Stanford University Press*, hal. 1–17.

[7] Xu, L., Xie, H., Qin, S. Z. J., Tao, X., & Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, hal. 1–20.

[8] Hu, E. J., et al. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, hal. 1–13.

[9] Wang, F., et al. (2025). A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2, hal. 6173–6183.

[10] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*.

[11] Guainazzo, N., Delzanno, G., Ancona, D., & D'Agostino, D. (2026). Navigating the seas of AI: Effectiveness of small language models on edge devices for maritime applications. *Sensors*, 26(5).

[12] Li, X., Lu, Z., Cai, D., Ma, X., & Xu, M. (2024). Large language models on mobile devices: Measurements, analysis, and insights. *EdgeFM 2024 - Proceedings of the 2024 Workshop on Edge and Mobile Foundation Models*, hal. 1–6.

[13] Wang, J., Zeng, Y., Guo, J., Ma, Y., Liu, A., & Liu, X. (2025). *SLMQuant: Benchmarking Small Language Model Quantization for Practical Deployment* (Vol. 1, No. 1). Association for Computing Machinery.

[14] Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., & Chen, J. (2025). A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8).

[15] Dutta, A., Ghosh, N., & Chatterjee, A. (2025). *CARE: A QLoRA-fine tuned multi-domain chatbot with fast learning on minimal hardware*.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

[16] Fan, J., Zhang, Y., Li, X., & Nikolopoulos, D. S. (2025). *Parallel CPU-GPU execution for LLM inference on constrained GPUs*.

[17] Hugging Face. (2026). Qwen2-1.5B-Instruct. Diambil dari https://huggingface.co/Qwen/Qwen2-1.5B-Instruct

[18] Blessing, L. T. M., & Chakrabarti, A. (2009). *DRM: A design research methodology*. Springer.

[19] Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publishing.

[20] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of EMNLP*, hal. 1–10.

[21] Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing*. Stanford University.

[22] Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization. *International Conference on Learning Representations (ICLR)*, hal. 1–20.

[23] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP*, hal. 3982–3992.

[24] Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge University Press.

[25] Liu, N. F., et al. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, hal. 157–173.

[26] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, hal. 633–642.

[27] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the Association for Computational Linguistics*, hal. 328–339.