Teknologi: Jurnal Ilmiah Sistem Informasi 16 (1) 2026 31-42

ISSN 2527-3671 (*Online*)
ISSN 2087-8893 (*Print*)

Tersedia *online* di www.journal.unipdu.ac.id

Unipdu

Halaman jurnal di www.journal.unipdu.ac.id/index.php/teknologi

Optimalisasi Chatbot Berbasis *TinyLLaMA* dengan *Fine - Tuning LoRA* pada Sistem Layanan Agensi Property

Fachrizal Fazza Ashari [a], *Agung Teguh Wibowo Almais* [b], *Fajar Rohman Hariri* [c]

[a,b,c] *Program Studi Teknik Informatika, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia*

email:[b,*] *agung.twa@ti.uin-malang.ac.id*

*Korespondensi

Dikirim 06 Mei 2026; **Direvisi** 18 Mei 2026; **Diterima** 15 Juni 2026; **Diterbitkan** 27 Juni 2026

Abstrak

Industri properti di Indonesia menghadapi tantangan dalam menyediakan layanan konsultasi yang cepat, responsif, dan efisien kepada calon pembeli. Model pemasaran tradisional dinilai kurang mampu memenuhi kebutuhan interaksi real-time yang semakin tinggi di era digital. Penelitian ini bertujuan mengembangkan chatbot konsultan properti berbahasa Indonesia menggunakan model TinyLLaMA-1.1B-Chat yang diadaptasi melalui metode fine-tuning Low-Rank Adaptation (LoRA) pada lapisan query projection (q_proj) dan value projection (v_proj). Dataset sintetis dikonstruksi menggunakan pendekatan Template-Based Data Generation yang mencakup tujuh kategori intent, yaitu simulasi KPR, dokumen legal, biaya dan pajak, spesifikasi properti, proses transaksi, konsultasi keputusan, serta risiko dan disclaimer. Eksperimen dilakukan pada tiga konfigurasi model dengan variasi nilai rank LoRA untuk menganalisis pengaruhnya terhadap kualitas dialog. Evaluasi dilakukan menggunakan metrik BLEU-4 serta kuesioner pengguna. BLEU-4 dipilih sebagai metrik kuantitatif utama karena kemampuannya mengukur kemiripan n-gram antara keluaran model dengan referensi secara objektif dan terukur, meskipun disadari bahwa metrik ini memiliki keterbatasan dalam menangkap variasi semantik pada tugas generasi dialog, sehingga dilengkapi dengan evaluasi pengguna sebagai bentuk validasi kualitatif yang dilakukan pada tahap uji implementasi model terbaik ke dalam sistem chatbot berbasis web. Hasil eksperimen menunjukkan bahwa Skenario 3 dengan konfigurasi rank r=24 dan lora_alpha=48 menghasilkan performa terbaik secara kuantitatif dengan skor BLEU-4 sebesar 16.57, mengungguli Skenario 1 (15.27) dan Skenario 2 (13.39). Hasil evaluasi pengguna mengindikasikan bahwa chatbot mampu menangani pertanyaan rutin secara efektif dengan performa yang sebanding dengan agen manusia, sehingga pendekatan hybrid antara chatbot dan agen manusia direkomendasikan untuk mengoptimalkan layanan secara menyeluruh.

Kata Kunci: TinyLLaMA, LoRA, Fine-Tuning, Chatbot, Properti

Optimizing a *TinyLLaMA-Based* Chatbot with *LoRA Fine-Tuning* for a Real Estate Agency Service System

Abstract

The property industry in Indonesia faces challenges in providing fast, responsive, and efficient consultation services to prospective buyers. Traditional marketing models are considered insufficient in meeting the growing demand for real-time interaction in the digital era. This study aims to develop an Indonesian-language property consultant chatbot using the TinyLLaMA-1.1B-Chat model, adapted through the Low-Rank Adaptation (LoRA) fine-tuning method applied to the query projection (q_proj) and value projection (v_proj) layers. A synthetic dataset was constructed using a Template-Based Data Generation approach, covering seven intent categories: mortgage simulation (KPR), legal documents, costs and taxes, property specifications, transaction processes, decision consultation, and risks and disclaimers. Experiments were conducted across three model configurations with varying LoRA rank values to analyze their effect on dialogue quality. Evaluation was performed using the BLEU-4 metric and a user questionnaire. BLEU-4 was selected as the primary quantitative metric due to its ability to objectively measure n-gram similarity between model outputs and reference answers, although it is acknowledged that this metric has limitations in capturing semantic variation in dialogue generation tasks, and was therefore complemented by user evaluation as a form of qualitative validation conducted during the implementation testing phase of the best-performing model in the web-based chatbot system. Experimental results show that Scenario 3, with a configuration of rank r=24 and lora_alpha=48, achieved the best quantitative performance with a BLEU-4 score of 16.57, outperforming Scenario 1 (15.27) and Scenario 2 (13.39). User evaluation results indicate that the chatbot is capable of handling routine inquiries effectively with performance comparable to human agents, thus a hybrid approach combining the chatbot and human agents is recommended to optimize the overall service experience.

Keywords: *: TinyLLaMA, LoRA, Fine-Tuning, Chatbot, Property*

Untuk mengutip artikel ini dengan *APA Style*:
Ashari, F. F., Almais, A. T. W., & Hariri, F. R. (2026). Optimalisasi Chatbot Berbasis TinyLLaMA dengan Fine - Tuning LoRA pada Sistem Layanan Agensi Property. TEKNOLOGI: Jurnal Ilmiah Sistem Informasi, 16(1), Halaman awal - Halaman akhir: https://doi.org/10.26594/teknologi.v16i1.6433

ISSN 2502-3357 (*Online*)
F.F. Ashari dkk./TEKNOLOGI 16(1) 2026 31-42
ISSN 2503-0477 (*Print*)

1. Pendahuluan

Transformasi digital telah mendorong perubahan mendasar dalam berbagai sektor ekonomi, termasuk industri properti. Sektor ini melibatkan transaksi bernilai tinggi yang menuntut kecepatan penyampaian informasi, jangkauan layanan yang luas, serta kemampuan merespons kebutuhan calon pembeli secara cepat dan akurat (Rahmaddion & Arribe, 2024). Namun, model layanan tradisional yang masih bergantung pada interaksi tatap muka, media cetak, dan pencatatan manual dinilai semakin tidak relevan dalam memenuhi ekspektasi konsumen modern yang menginginkan akses informasi secara instan (Faturahman et al., 2023). Di Indonesia, kondisi ini semakin terasa seiring pesatnya urbanisasi dan meningkatnya kebutuhan hunian, yang mendorong permintaan akan sistem layanan properti yang lebih adaptif dan responsif.

Upaya digitalisasi melalui sistem informasi berbasis web telah membawa kemajuan dalam pengelolaan data properti, namun belum sepenuhnya menjawab kebutuhan interaksi pelanggan secara real-time (Rahmaddion & Arribe, 2024). Pelanggan modern tidak hanya membutuhkan informasi yang statis, melainkan juga konsultasi dinamis yang mampu menjawab pertanyaan spesifik, seperti simulasi cicilan KPR, rekomendasi properti sesuai kondisi finansial, hingga penjelasan proses transaksi secara bertahap. Kesenjangan antara kebutuhan pasar tersebut dengan kapabilitas sistem yang ada menegaskan perlunya solusi berbasis kecerdasan buatan yang mampu menghadirkan interaksi percakapan secara otomatis dan kontekstual.

Teknologi Large Language Models (LLM) menawarkan potensi besar dalam menjawab tantangan tersebut. LLM seperti LLaMA mampu menghasilkan teks alami yang menyerupai respons manusia, sehingga dapat meningkatkan kualitas komunikasi antara agen dan pelanggan secara lebih efisien (Herlawati & Handayanto, 2026). Meski demikian, penerapan LLM pada domain spesifik seperti industri properti masih menghadapi keterbatasan signifikan. Model LLM yang dilatih pada data umum cenderung menghasilkan respons yang tidak sesuai dengan terminologi, gaya tutur, dan konteks spesifik domain properti berbahasa Indonesia (Yu & McGuinness, 2024). Proses adaptasi melalui fine-tuning pada data domain-spesifik menjadi solusi yang diperlukan, namun pendekatan full fine-tuning membutuhkan sumber daya komputasi yang sangat besar dan tidak selalu memungkinkan dalam kondisi infrastruktur terbatas.

Metode Low-Rank Adaptation (LoRA) hadir sebagai solusi parameter-efficient fine-tuning yang efektif dan efisien secara komputasi. LoRA bekerja dengan hanya melatih matriks low-rank tambahan pada lapisan tertentu, tanpa memperbarui seluruh parameter model dasar, sehingga mampu menurunkan kebutuhan memori dan waktu pelatihan secara signifikan tanpa mengorbankan kualitas hasil adaptasi (Huan & Shun, 2025). Sejumlah studi telah membuktikan bahwa LoRA mampu mempertahankan performa model yang kompetitif bahkan pada skenario keterbatasan perangkat keras (Hu et al., 2021). Dengan demikian, LoRA menjadi pilihan yang relevan untuk mengadaptasi LLM pada domain spesifik dengan anggaran komputasi terbatas.

Dalam penelitian ini, TinyLLaMA-1.1B-Chat dipilih sebagai model dasar dengan mempertimbangkan ukurannya yang ringan, yakni hanya 1.1 miliar parameter, sehingga cocok untuk implementasi pada perangkat keras yang tidak memerlukan infrastruktur GPU kelas tinggi (Zhang et al., 2024). Meskipun berukuran kecil, TinyLLaMA telah menunjukkan performa yang kompetitif pada berbagai tugas percakapan dan terbukti dapat ditingkatkan secara signifikan melalui fine-tuning pada data domain-spesifik.

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi chatbot konsultan properti berbahasa Indonesia dengan melakukan fine-tuning LoRA pada model TinyLLaMA-1.1B-Chat. Dataset sintetis dikonstruksi menggunakan pendekatan Template-Based Data Generation yang mencakup tujuh kategori intent layanan properti. Eksperimen dilakukan pada empat konfigurasi model dengan variasi nilai rank LoRA pada lapisan query projection (q_proj) dan value projection (v_proj) untuk menganalisis pengaruhnya terhadap kualitas perilaku dialog. Evaluasi dilakukan menggunakan metrik BLEU-4 serta kuesioner pengguna untuk mengukur efektivitas sistem dari perspektif praktisi. Hasil penelitian ini diharapkan dapat menjadi referensi pengembangan chatbot domain-spesifik berbahasa Indonesia menggunakan model LLM yang ringan dan efisien.

2. *State of the Art*

2.1. *Penelitian Terkait*

Sejumlah penelitian telah dilakukan pada bidang fine-tuning LLM dan pengembangan chatbot domain spesifik. Meyer et al. (2024) membandingkan tiga pendekatan fine-tuning utama, yaitu QLoRA, RAFT, dan RLHF. Hasil studi tersebut menunjukkan bahwa QLoRA unggul dari sisi efisiensi penggunaan memori, RAFT lebih baik dalam menjaga struktur keluaran, sementara RLHF menghasilkan respons yang lebih natural karena berbasis preferensi manusia. Perbandingan ini memberikan gambaran komprehensif mengenai kelebihan dan keterbatasan masing-masing metode dalam konteks pengembangan chatbot.

Dalam hal efisiensi adaptasi model, Hu et al. (2021) memperkenalkan LoRA sebagai metode parameter-efficient fine-tuning yang membekukan bobot utama model dan hanya melatih matriks low-rank tambahan pada lapisan tertentu. Pendekatan ini terbukti mampu menurunkan kebutuhan memori secara signifikan tanpa mengorbankan performa model. Pengembangan lebih lanjut dilakukan oleh Liu et al. (2025) melalui RoRA yang menambahkan regularisasi rank untuk meningkatkan stabilitas pelatihan, serta Dettmers et al. (2023) melalui QLoRA yang menggabungkan LoRA dengan teknik kuantisasi 4-bit sehingga memungkinkan fine-tuning model hingga 65 miliar parameter menggunakan GPU kelas menengah.

Terkait penelitian, Zhang et al. (2024) memperkenalkan TinyLLaMA, sebuah model open source dengan 1.1 miliar parameter yang mampu menjalankan tugas percakapan dengan performa kompetitif. Studi Fu et al. (2024) melalui penelitiannya mengonfirmasi bahwa LLM berparameter kecil mampu menghasilkan ringkasan rapat yang mendekati kualitas model besar dengan biaya komputasi jauh lebih rendah. Sementara Wu et al. (2024) melalui model GEB menunjukkan bahwa LLM ringan juga dapat dikembangkan untuk mendukung kemampuan multibahasa. Studi-studi ini menegaskan bahwa model kecil tetap relevan dan kompetitif apabila proses pelatihan dan adaptasinya dilakukan secara optimal.

Dari sisi aplikasi chatbot, Pratap et al. (2025) melalui tinjauan literatur menyeluruh menemukan bahwa keberhasilan chatbot tidak hanya ditentukan oleh algoritma, tetapi juga oleh penerimaan pengguna, integrasi sistem, dan konteks aplikasinya. Y. Zhang et al. (2024) dalam studi sistematis mengenai chatbot bisnis berbasis deep learning memetakan taksonomi aplikasi dan tren riset yang relevan, sementara Pandya & Mahavidyalaya (2023) membuktikan potensi integrasi LLM secara langsung dalam sistem layanan pelanggan menggunakan LangChain dan model Flan-T5. Secara keseluruhan, berbagai penelitian tersebut menunjukkan adanya sinergi antara inovasi teknis PEFT, pengembangan model ringan, dan penerapan praktis chatbot di berbagai domain layanan.

2.2. *Template-Based Data Generation*

Template-Based Data Generation merupakan pendekatan yang efektif dalam mengonstruksi dataset sintetis untuk pelatihan model. Metode ini dimulai dari serangkaian templat pertanyaan dasar yang memiliki placeholder yang dapat divariasikan untuk menghasilkan beragam pasangan instruksi dan respons secara otomatis tanpa membutuhkan anotasi manual dalam jumlah besar (Wang et al., 2023). Dalam konteks chatbot, pendekatan ini memungkinkan pembuatan dataset yang mencakup berbagai profil pengguna dan skenario percakapan, sehingga model yang dilatih mampu menangani pertanyaan dari berbagai konteks secara lebih adaptif (Ding et al., 2023).

Meski demikian, pendekatan ini memiliki keterbatasan dalam hal representasi data untuk tugas yang lebih kompleks. Maheshwari et al. (2024) menunjukkan bahwa data sintetis hasil generasi templat dapat kurang representatif untuk tugas seperti Named Entity Recognition (NER) yang memerlukan pemahaman konteks mendalam. Oleh karena itu, diperlukan proses penyaringan dan validasi kualitas dataset secara cermat sebelum digunakan dalam proses fine-tuning, termasuk evaluasi oleh evaluator berlatar belakang domain terkait (Paula et al., 2021).

2.3. *Model TinyLLaMA*

TinyLLaMA merupakan model bahasa open-source berbasis arsitektur decoder-only Transformer dengan jumlah parameter sekitar 1.1 miliar. Model ini dirancang untuk menjawab kebutuhan akan LLM yang ringan dan dapat dijalankan pada perangkat dengan keterbatasan sumber daya, seperti laptop ber-GPU menengah atau server dengan kapasitas terbatas (Zhang et al., 2024). Meskipun berukuran kecil, TinyLLaMA telah menjalani proses pretraining pada dataset yang cukup besar sehingga menghasilkan kemampuan pemahaman bahasa alami yang relevan untuk berbagai tugas percakapan, klasifikasi teks, dan question answering. Karakteristik ini menjadikannya pilihan yang praktis untuk implementasi chatbot domain spesifik, terutama pada lingkungan dengan keterbatasan infrastruktur komputasi.

2.4. *Tantangan Pengolahan Data Bahasa pada Model Kecil (1.1B)*

Evaluasi kemampuan LLM dalam memahami bahasa Indonesia mengungkapkan kesenjangan yang signifikan antara model berukuran besar dan kecil, khususnya pada pemrosesan konten linguistik dan

ISSN 2502-3357 (*Online*)
F.F. Ashari dkk./TEKNOLOGI 16(1) 2026 31-42
ISSN 2503-0477 (*Print*)

budaya lokal. Koto et al. (2023) melalui benchmark IndoMMLU menunjukkan bahwa GPT-3.5 dengan 175 miliar parameter hanya mampu melewati ujian setingkat sekolah dasar di Indonesia dengan akurasi keseluruhan 53.2%, sementara model-model kecil seperti BLOOMZ (560M–7.1B) dan XGLM (564M–7.5B) bahkan kesulitan melampaui performa acak pada sebagian besar kategori soal. Lebih jauh, model dengan parameter sekitar 1.1 miliar seperti BLOOMZ (1.1B) hanya mencapai akurasi rata-rata 22.4%, yang hampir setara dengan baseline acak sebesar 24.4%, mengindikasikan bahwa model sekelas TinyLLaMA (1.1B) menghadapi tantangan mendasar dalam memahami struktur bahasa Indonesia yang kompleks, termasuk konstruksi negasi yang umum digunakan dalam konteks pendidikan formal maupun konten budaya lokal yang sangat minim representasinya dalam data pretraining model-model tersebut.

2.5. Metode Low-Rank Adaptation (LoRA)

LoRA merupakan metode parameter-efficient fine-tuning yang bekerja dengan membekukan sebagian besar bobot utama model pra-latih, kemudian menyisipkan matriks low-rank yang dapat dilatih pada lapisan tertentu (Hu et al., 2021). Secara matematis, pembaruan bobot pada LoRA direpresentasikan sebagai $\boldsymbol{\Delta W} = \boldsymbol{A} \times \boldsymbol{B}$, di mana $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ dan $\boldsymbol{B} \in \mathbb{R}^{r \times k}$ adalah matriks adaptasi dengan rank $\boldsymbol{r} \ll \mathbf{min}(\boldsymbol{d}, \boldsymbol{k})$. Jumlah parameter yang dilatih menjadi $\boldsymbol{r} \times (\boldsymbol{d} + \boldsymbol{k})$, jauh lebih kecil dibandingkan full fine-tuning yang memerlukan d\times k parameter. Pendekatan ini terbukti mampu menurunkan kebutuhan memori secara signifikan tanpa mengurangi kualitas adaptasi model terhadap domain spesifik (Huan & Shun, 2025).

Dalam penelitian ini, LoRA diterapkan secara spesifik pada lapisan query projection (q_proj) dan value projection (v_proj) dari layer self-attention TinyLLaMA. Variasi nilai rank pada ketiga konfigurasi model digunakan untuk menganalisis pengaruh kapasitas representasi terhadap kualitas perilaku dialog chatbot yang dihasilkan.

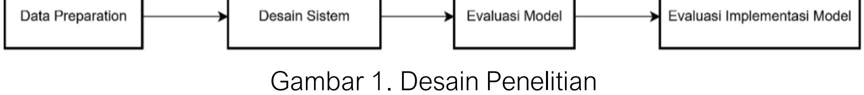
2.6. Metrik Evaluasi

Evaluasi kualitas keluaran model dilakukan menggunakan metrik BLEU (Bilingual Evaluation Understudy), yaitu metrik berbasis precision yang menilai tingkat kemiripan n-gram antara teks yang dihasilkan model dengan teks referensi. BLEU dilengkapi dengan brevity penalty untuk mencegah bias terhadap jawaban yang terlalu pendek, sehingga model didorong untuk menghasilkan respons yang tidak hanya tepat secara frasa tetapi juga memadai secara panjang (Yang et al., 2018). Semakin tinggi skor BLEU yang diperoleh, semakin besar tingkat kemiripan keluaran model dengan jawaban referensi yang telah ditetapkan.

3. Metode Penelitian

3.1. Desain Penelitian

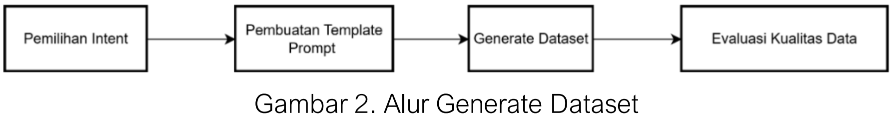
Penelitian ini dirancang dalam empat tahap utama yang saling berkesinambungan, sebagaimana diilustrasikan pada Gambar 1. Tahap pertama adalah data preparation, yang mencakup pemilihan intent, penentuan format dataset, konstruksi template prompt, dan proses generasi data sintetis menggunakan pendekatan Template-Based Data Generation. Tahap kedua adalah desain sistem, meliputi pemilihan model dasar TinyLLaMA-1.1B-Chat-v1.0, penentuan konfigurasi LoRA, serta penetapan hiperparameter pelatihan. Tahap ketiga adalah pelatihan dan evaluasi model, di mana empat konfigurasi model dilatih dan dibandingkan menggunakan metrik BLEU serta kuesioner pengguna. Tahap keempat adalah uji implementasi, yaitu pengujian chatbot secara langsung dengan membandingkan performa chatbot terhadap agen manusia melalui penilaian responden.



Gambar 1. Desain Penelitian

3.2. Data Preparation

Dataset yang digunakan dalam penelitian ini bersifat sintetis dan dikonstruksi menggunakan pendekatan *Template-Based Data Generation*. Proses konstruksi dataset terdiri dari empat tahap berurutan sebagaimana ditunjukkan pada Gambar 2, yaitu pemilihan *intent*, pembuatan *template prompt*, *generate dataset*, dan evaluasi kualitas data.



Gambar 2. Alur Generate Dataset

Proses diawali dengan pemilihan tujuh kategori *intent* yang merepresentasikan topik percakapan paling relevan dalam layanan konsultasi properti. Setiap kategori intent dirancang untuk mencakup variasi pertanyaan dari berbagai profil pengguna, meliputi first-time buyer, pasangan, investor, maupun pembeli tunggal. Deskripsi lengkap ketujuh kategori intent disajikan pada Tabel 1.

Tabel 1. kategori Intent

Intent	Deskripsi
kpr_simulasi	Simulasi perhitungan kemampuan finansial konsumen dalam membayar cicilan KPR jangka panjang (Akhsya et al., 2021).
legal_dokumen	Verifikasi legalitas transaksi properti meliputi sertifikat tanah, IMB, dan perjanjian jual beli (Septiana et al., 2020).
biaya_dan_pajak	Informasi biaya tambahan pembelian properti seperti pajak, biaya notaris, dan administrasi (Septiana et al., 2020).
spesifikasi_properti	Detail ukuran, desain, kualitas bangunan, dan fasilitas properti sesuai kebutuhan konsumen (Akhsya et al., 2021).
proses_transaksi	Tahapan pembelian rumah meliputi penentuan harga, pembayaran, dan penandatanganan dokumen legal (Septiana et al., 2020).
konsultasi_keputusan	Panduan konsultasi pemilihan properti dari sisi harga, lokasi, dan legalitas transaksi (Akhsya et al., 2021).
risiko_dan_disclaimer	Informasi potensi risiko pembelian properti terkait hukum, kualitas bangunan, dan pengelolaan (Akhsya et al., 2021).

Selanjutnya, tahap pembuatan template prompt dirancang dengan aturan ketat agar data yang dihasilkan berkualitas tinggi, di antaranya setiap data harus unik dalam struktur kalimat dan skenario, jawaban minimal 120 kata, mengandung penjelasan bertahap atau poin bernomor, serta dilarang mengandung konten hard selling. Format dataset mengikuti struktur instruction-output dengan empat atribut, yaitu intent (kategori tujuan percakapan), instruction (pertanyaan alami dari calon pembeli), input (konteks tambahan yang bersifat opsional), dan output (jawaban analitis dan terstruktur dari chatbot). Contoh format dataset disajikan pada Tabel 2.

Tabel 2. Contoh Dataset

Atribut	Tipe Data	Contoh
intent	String	“kpr_simulasi”
instruction	String	" Saya karyawan swasta dengan gaji tetap Rp8,7 juta per bulan dan belum pernah ambil KPR. Kalau saya mau beli rumah seharga Rp325 juta, kira-kira cicilannya aman atau tidak ya?"
input	String	""
output	String	"Baik, saya bantu dengan simulasi sederhana agar Anda punya gambaran awal. Asumsi yang saya gunakan: harga rumah Rp325 juta, uang muka 15%, tenor 20 tahun, dan bunga efektif rata-rata 9% per tahun.\n \n1) Perhitungan DP: 15% dari harga berarti sekitar Rp48.75 juta di awal, belum termasuk biaya lain. \n2) Plafon pinjaman: Sisa yang dibiayai bank sekitar Rp276 jutaan. \n3) Estimasi cicilan: Dengan tenor 20 tahun dan bunga kisaran tersebut, cicilan bulanan bisa berada di rentang Rp2.4–Rp2.7 juta. \n4) Rasio terhadap gaji: Dengan gaji Rp8.7 juta, cicilan ini sekitar 28–31% dari penghasilan, masih dalam batas umum perbankan.\n \nNamun perlu diperhitungkan juga utang lain, biaya hidup rutin, dan dana darurat. Simulasi ini bersifat estimasi, karena suku bunga dan kebijakan bank bisa berbeda.”

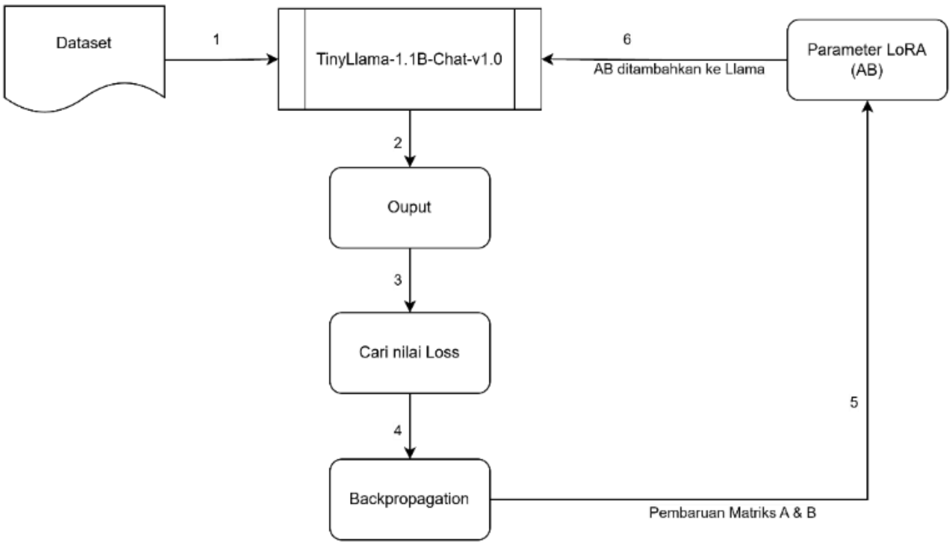
ISSN 2502-3357 (*Online*)
F.F. Ashari dkk./TEKNOLOGI 16(1) 2026 31-42
ISSN 2503-0477 (*Print*)

Pada tahap generate dataset, proses generasi dilakukan menggunakan model GPT dengan target 3.042 data percakapan yang mencakup seluruh kategori intent secara proporsional. Dataset yang telah terkumpul kemudian memasuki tahap evaluasi kualitas data sebelum dinyatakan layak digunakan, dengan proporsi pembagian 95% untuk pelatihan dan 5% untuk validasi, serta pembatasan sekitar 300 data untuk set evaluasi guna memastikan penilaian yang representatif.

Untuk memastikan kualitas dataset sintetis yang dihasilkan, dilakukan evaluasi oleh 5 evaluator berlatar belakang domain properti melalui kuesioner terstruktur yang mencakup dua aspek utama, yaitu identifikasi intent dan variasi jawaban, serta penilaian realitas dan kelayakan jawaban. Dataset terlebih dahulu dibagi menjadi subset berdasarkan tujuh kategori intent, di mana setiap kategori diwakili oleh 2–3 data representatif untuk dinilai. Evaluator diminta menilai keberagaman pertanyaan dalam konteks profil pengguna yang bervariasi, sekaligus mengevaluasi apakah jawaban chatbot terasa alami, mengandung informasi yang relevan, dan mudah dipahami oleh calon pembeli properti. Data yang dinilai kurang relevan atau tidak akurat berdasarkan hasil kuesioner kemudian diperbaiki sebelum dataset dinyatakan layak digunakan dalam proses fine-tuning. Pendekatan evaluasi berbasis domain expert ini dipilih sebagai pengganti metrik kesepakatan antar-anotator formal seperti Cohen's Kappa, mengingat sifat dataset yang bersifat domain-spesifik dan memerlukan pertimbangan kontekstual dari praktisi properti secara langsung.

3.3. Desain Sistem

Desain sistem pada penelitian ini memetakan alur kerja secara menyeluruh mulai dari dataset hingga menghasilkan model chatbot yang siap digunakan, sebagaimana diilustrasikan pada Gambar 3. Alur sistem dimulai dari dataset sintetis yang terlebih dahulu dibagi menjadi dua bagian, yaitu *training data* dan *evaluation data*. Data pelatihan kemudian digunakan dalam proses fine-tuning model TinyLLaMA-1.1B-Chat-v1.0 menggunakan pendekatan LoRA.



Gambar 3. Desain Sistem

A. *Inisialisasi Parameter LoRA*

Pada tahap inisialisasi LoRA, dua matriks tambahan ditentukan, yaitu matriks $\boldsymbol{A} \in \mathbb{R}^{r\times d\_model}$ dan matriks $\boldsymbol{B} \in \mathbb{R}^{d\_model\times r}$, di mana $\boldsymbol{r}$ dinyatakan sebagai nilai *rank* dan *d_model* adalah dimensi model. Bobot asli model ditulis sebagai $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ dan tetap dibekukan, sementara kedua matriks A dan B dilatih untuk membentuk representasi *low rank* dari pembaruan parameter. Output akhir layer setelah penyesuaian LoRA dihitung sebagai:

$$\boldsymbol{W}' = \boldsymbol{W} + (\boldsymbol{A} \cdot \boldsymbol{B})$$

Disini matriks $\boldsymbol{A} \in \mathbb{R}^{r\times d}$ dianggap sebagai matriks *down rank projection*, sementara $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ dianggap sebagai matriks *up rank projection*, dan $\boldsymbol{AB} \in \mathbb{R}^{d\times k}$ merupakan matriks *update low rank* yang dilatih selama proses *fine-tuning* berjalan. Dalam penelitian ini, parameter LoRA ditambahkan pada proyeksi $\boldsymbol{W_Q}$ atau $\boldsymbol{W_V}$, sehingga bobot efektif yang digunakan menjadi $\boldsymbol{W} + \boldsymbol{AB}$. Pendekatan ini memungkinkan adaptasi model secara efisien tanpa mengubah seluruh bobot asli TinyLLaMA-1.1B-Chat-v1.0.

B. Proses Pelatihan

Sebagaimana diilustrasikan pada Gambar 3, selama proses pelatihan data diinput dan diproses oleh model untuk menghasilkan *output*. Nilai *loss* kemudian dihitung menggunakan fungsi *cross-entropy* untuk mengukur kesalahan prediksi token:

$$\boldsymbol{L = -\log(p_y)}$$

di mana $\boldsymbol{T}$ adalah panjang sekuens,$\boldsymbol{x_t}$ adalah token target pada posisi ke-t, dan $\boldsymbol{P(\ x_t \mid x_{<t})}$ adalah probabilitas token yang diprediksi model. Melalui *backpropagation*, gradien dihitung terhadap parameter A dan B:

$$\boldsymbol{\frac{\partial L}{\partial A}, \frac{\partial L}{\partial B}}$$

Digunakan untuk memperbarui hanya matriks A dan B, sementara seluruh bobot asli model $\boldsymbol{W_0}$ tetap dibekukan. Proses ini diulang hingga nilai *loss* konvergen, setelah itu parameter BA yang telah diperbarui dijumlahkan ke dalam bobot model dasar TinyLLaMA.

C. Arsitektur Transformer dan Integrasi LoRA

Gambar 3 mengilustrasikan TinyLLaMA beserta integrasi parameter LoRA. Alur pemrosesan dimulai dari *input* teks yang melalui tahap *tokenization*, kemudian diproyeksikan ke ruang embedding berdimensi tinggi. Representasi embedding tersebut selanjutnya diproses secara sekuensial melalui N blok *decoder*, di mana setiap blok mengandung mekanisme *self-attention* dan *feed-forward network*. Keluaran akhir dari *decoder* ke-N menghasilkan *output* berupa distribusi probabilitas token berikutnya.

Mekanisme self-attention pada setiap blok decoder diformulasikan sebagai:

$$\boldsymbol{Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V}$$

di mana Q, K, dan V masing-masing merupakan matriks *query*, *key*, dan *value* yang diperoleh melalui proyeksi linear dari embedding X:

$$\boldsymbol{Q = X.W_Q, \qquad K = X.W_K, \qquad V = X.W_V}$$

Selanjutnya, $\boldsymbol{d_k}$ adalah dimensi vektor *key* yang berfungsi sebagai faktor penskalaan untuk menstabilkan gradien. Integrasi LoRA pada penelitian ini diterapkan khusus pada lapisan proyeksi *query* $(\boldsymbol{W_Q})$ dan *value* $(\boldsymbol{W_V})$, sehingga bobot efektif kedua proyeksi tersebut menjadi:

$$\boldsymbol{Q = X(W_Q + B_Q A_Q), \qquad V = X(W_V + B_V A_V)}$$

di mana $(\boldsymbol{A_{Q'} B_Q})$ dan $(\boldsymbol{A_{V'} B_V})$ adalah pasangan matriks LoRA yang dilatih masing-masing untuk proyeksi *query* dan *value*, sementara $\boldsymbol{W_K}$ tidak dimodifikasi. Pendekatan ini memungkinkan model mempelajari pola perilaku dialog domain properti hanya melalui matriks berdimensi rendah tanpa mengubah seluruh bobot pretrained, sehingga kebutuhan komputasi selama *fine-tuning* dapat diminimalkan secara signifikan.

3.4. Skenario Uji

Proses *fine-tuning* dilakukan pada model dasar TinyLLaMA-1.1B-Chat-v1.0 menggunakan metode LoRA yang disisipkan pada lapisan *query projection* (q_proj) dan *value projection* (v_proj) dari mekanisme *self-attention*. Pemilihan kedua lapisan ini didasarkan pada pertimbangan bahwa sebagian besar kapasitas generalisasi model terletak pada blok perhatian, sehingga mengadaptasi bagian ini sudah cukup untuk membentuk perilaku dialog domain properti tanpa memperbarui seluruh bobot model. Seluruh proses pelatihan menggunakan optimizer AdamW dengan *learning rate scheduler* bertipe kosinus, serta menerapkan mekanisme *early stopping* untuk mencegah *overfitting*.

Eksperimen dilakukan pada tiga konfigurasi model dengan variasi hiperparameter LoRA sebagaimana ditunjukkan pada Tabel 3. Skenario 1 merupakan konfigurasi *fine-tuning* awal dengan *rank* 16 dan *learning rate* 2e-4. Skenario 2 menggunakan rank yang sama namun dengan *dropout* lebih besar, *learning rate* lebih rendah, dan jumlah *epoch* lebih banyak untuk mendorong regularisasi yang lebih ketat. Skenario 3 dirancang dengan konfigurasi paling agresif menggunakan rank 24 dan *learning rate* 2.2e-4 untuk menguji pengaruh kapasitas representasi yang lebih tinggi terhadap kualitas dialog.

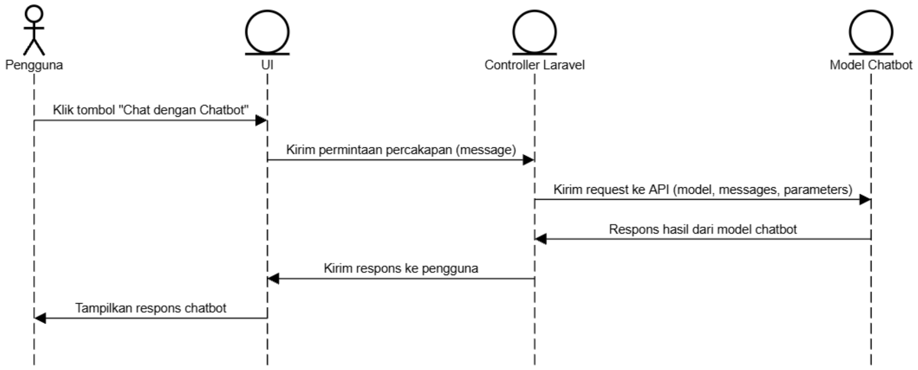
Tabel 3. Skenario Uji

Skenario	LoRA Rank (r)	LoRA Alpha	LoRA Dropout	Learning Rate	Epoch
Skenario 1	16	32	0.05	2e-4	8
Skenario 2	16	32	0.1	1.5e-4	10
Skenario 3	24	48	0.05	2.2e-4	12

ISSN 2502-3357 (*Online*)
F.F. Ashari dkk./TEKNOLOGI 16(1) 2026 31-42
ISSN 2503-0477 (*Print*)

3.5. Implementasi Model

Implementasi chatbot menggunakan framework Laravel dengan *ilab* sebagai server yang menjalankan LLM. Model hasil *fine-tuning* kemudian dikonversi dari format Safetensor ke format GGUF menggunakan library *llama-cpp*, kemudian dimuat ke dalam server *ilab* agar siap menerima permintaan melalui Chatbot API.



Gambar 4. Sequence Diagram Implementasi Model terhadap Sistem

Alur komunikasi sistem ditunjukkan pada Gambar 4. Pengguna mengirimkan pesan melalui antarmuka *frontend*, yang kemudian diteruskan via HTTP POST pada backend Laravel. Backend meneruskan permintaan tersebut ke Chatbot API untuk diproses, dan respons dikembalikan ke *frontend* secara bertahap menggunakan mekanisme *Server-Sent Events (SSE)*. Pendekatan *streaming* ini memungkinkan token respons ditampilkan secara dinamis tanpa perlu penyegaran halaman, sehingga menghasilkan pengalaman interaksi yang lebih responsif bagi pengguna.

3.6. Evaluasi Model

Evaluasi kualitas keluaran model dilakukan menggunakan metrik BLEU-4 terhadap 300 data validasi menggunakan pustaka *evaluate* dari Hugging Face. Skor BLEU-4 dihitung sebagai berikut:

$$\boldsymbol{BLEU = BP \times \left(\sum_{n=1}^{N} \omega_n \log p_n\right)}$$

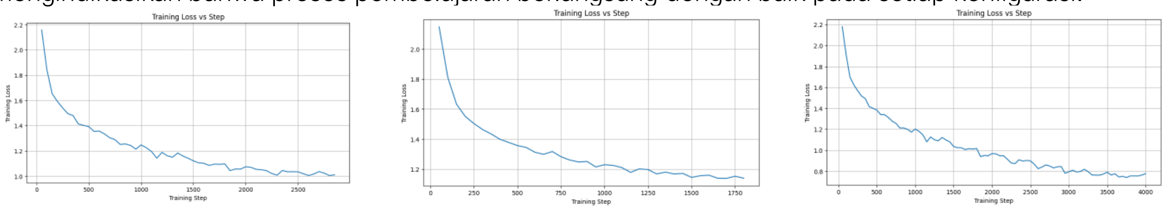
Brevity penalty, $\boldsymbol{p_n}$ adalah presisi *n-gram* ke-n, dan $\boldsymbol{\omega_n}$ adalah bobotnya. Skor yang lebih tinggi mengindikasikan kesesuaian keluaran model yang lebih baik terhadap jawaban referensi. Selain evaluasi kuantitatif, dilakukan uji implementasi melalui kuesioner pengguna yang mengukur kecepatan respons dan kualitas jawaban chatbot dibandingkan agen manusia pada tujuh topik layanan properti.

4. Hasil dan Pembahasan

4.1 Hasil Pelatihan Model

Proses fine-tuning dilakukan pada tiga skenario konfigurasi model menggunakan pendekatan LoRA pada lapisan q_proj dan v_proj model TinyLLaMA-1.1B-Chat-v1.0. Seluruh proses pelatihan menggunakan GPU dengan pembagian data 95% untuk pelatihan dan 5% untuk validasi. Mekanisme early stopping diterapkan untuk menghentikan pelatihan secara otomatis apabila tidak terjadi peningkatan performa pada data validasi dalam beberapa evaluasi berturut-turut, guna mencegah overfitting sekaligus menemukan konfigurasi yang paling stabil.

Gambar 5 menyajikan kurva Training Loss vs Step dari ketiga Skenario Uji. Secara umum, ketiga skenario menunjukkan tren penurunan loss yang konsisten dari awal hingga akhir pelatihan, mengindikasikan bahwa proses pembelajaran berlangsung dengan baik pada setiap konfigurasi.



(a) (b) (c)

Gambar 5. Hasil Fine-Tuning loRA: (a) Skenario Uji 1; (b) Skenario Uji 2; (c) Skenario Uji 3

Skenario 1 dilatih dengan learning rate 2e-4 dan LoRA rank 16 selama 8 epoch. Sebagaimana terlihat pada Gambar 5(a), kurva loss mengalami penurunan yang relatif cepat di awal pelatihan namun disertai fluktuasi yang cukup terlihat sepanjang proses, dengan nilai loss akhir yang stabil di kisaran 1.0–1.2 pada sekitar 2.700 training step. Skenario 2 dikembangkan sebagai konfigurasi yang lebih stabil dengan

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)
menurunkan learning rate menjadi 1.5e-4, meningkatkan dropout menjadi 0.1, dan menambah jumlah epoch menjadi 10. Pada Gambar 5(b), kurva loss menunjukkan pola yang lebih halus dan konsisten dibandingkan Skenario 1, dengan fluktuasi yang lebih kecil dan konvergensi yang lebih mulus hingga sekitar 1.750 training step, mencerminkan efek regularisasi yang lebih ketat dari konfigurasi ini. Skenario 3 dirancang dengan kapasitas adaptasi tertinggi melalui peningkatan rank LoRA menjadi 24 dan alpha menjadi 48 dengan learning rate 2.2e-4 dan 12 epoch. Gambar 5(c) memperlihatkan bahwa meskipun fluktuasi masih terjadi pada fase awal, kurva loss secara keseluruhan menunjukkan penurunan yang lebih dalam dan berkelanjutan hingga mendekati nilai 0.8 pada sekitar 4.000 training step. Hal ini mengindikasikan konvergensi yang lebih mendalam terhadap pola domain properti dibandingkan kedua skenario sebelumnya, sebagaimana tercermin pada penurunan average training loss per epoch yang lebih konsisten hingga epoch terakhir.

4.2 Evaluasi Model

Evaluasi kuantitatif dilakukan menggunakan metrik BLEU-4 data validasi yang telah dipisahkan dari dataset utama menggunakan pustaka evaluate dari Hugging Face. Perbandingan skor BLEU-4 ketiga skenario disajikan pada Tabel 4.

Tabel 4. Tabel Hasil Evaluasi Metrik BLEU-4

Skenario	Skor BLEU-4
Skenario 1	15.27
Skenario 2	13.39
Skenario 3	16.57

Hasil evaluasi menunjukkan bahwa Skenario 3 menghasilkan skor BLEU-4 tertinggi sebesar 16.57, diikuti Skenario 1 dengan skor 15.27, dan Skenario 2 dengan skor 13.39. Skenario 3 yang menggunakan rank LoRA lebih besar yaitu r=24 dengan alpha=48 terbukti menghasilkan kapasitas representasi yang lebih tinggi, sehingga model mampu mempelajari pola frasa dan struktur jawaban domain properti secara lebih mendalam dibandingkan kedua skenario lainnya. Hal ini sejalan dengan temuan Hu et al. (2021) yang menunjukkan bahwa nilai rank yang lebih besar pada adaptor LoRA meningkatkan fleksibilitas model dalam mengadaptasi gaya dan struktur keluaran pada domain spesifik.

Menariknya, Skenario 2 menghasilkan skor BLEU-4 terendah meskipun menggunakan jumlah epoch lebih banyak dibandingkan Skenario 1. Hal ini mengindikasikan bahwa kombinasi learning rate yang lebih rendah (1.5e-4) dengan dropout yang lebih besar (0.1) pada Skenario 2 mendorong regularisasi yang terlalu ketat, sehingga membatasi kemampuan model dalam menyesuaikan pola keluaran terhadap data referensi. Temuan ini menunjukkan bahwa peningkatan kapasitas representasi melalui rank yang lebih besar, sebagaimana diterapkan pada Skenario 3, lebih efektif dalam meningkatkan kualitas dialog dibandingkan sekadar memperpanjang durasi pelatihan dengan regularisasi yang lebih kuat.

Penggunaan metriks BLEU pada tugas generasi dialog cenderung menghasilkan skor yang lebih rendah dibandingkan pada mesin terjemahan, mengingat satu pertanyaan dapat memiliki banyak jawaban yang valid secara semantik namun berbeda secara leksikal dari referensi yang ditetapkan, sehingga model generatif kerap mendapat penalti meskipun jawabannya kontekstual tepat. Konteks ini perlu dipertimbangkan ketika menginterpretasikan hasil penelitian, mengingat TinyLLaMA dengan kapasitas 1.1 miliar parameter memiliki ruang representasi yang jauh lebih terbatas dibandingkan model besar pada umumnya. Dengan demikian, pencapaian skor BLEU-4 pada rentang 13–16% untuk tugas generasi dialog domain spesifik berbahasa Indonesia menggunakan model sekecil ini sudah mencerminkan adanya progres adaptasi terhadap pengetahuan baru walaupun masih kurang optimal melalui fine-tuning LoRA, meskipun eksplorasi metrik evaluasi tambahan seperti ROUGE-L atau BERTScore pada penelitian selanjutnya disarankan untuk memberikan gambaran yang lebih komprehensif terhadap kualitas keluaran model.

4.3 Implementasi Sistem Chatbot

Model Skenario 3 sebagai model terbaik hasil evaluasi BLEU diintegrasikan ke dalam sistem chatbot berbasis web menggunakan framework Laravel dengan ilab sebagai server yang menjalankan LLM. Model hasil fine-tuning dikonversi dari format Safetensor ke format GGUF menggunakan library llama-cpp, kemudian dimuat ke dalam server ilab agar siap menerima permintaan melalui Chatbot API. Sistem dijalankan pada perangkat dengan spesifikasi Intel Core i5 Gen 13420H, GPU RTX 4050 VRAM 6GB, dan RAM 8GB.

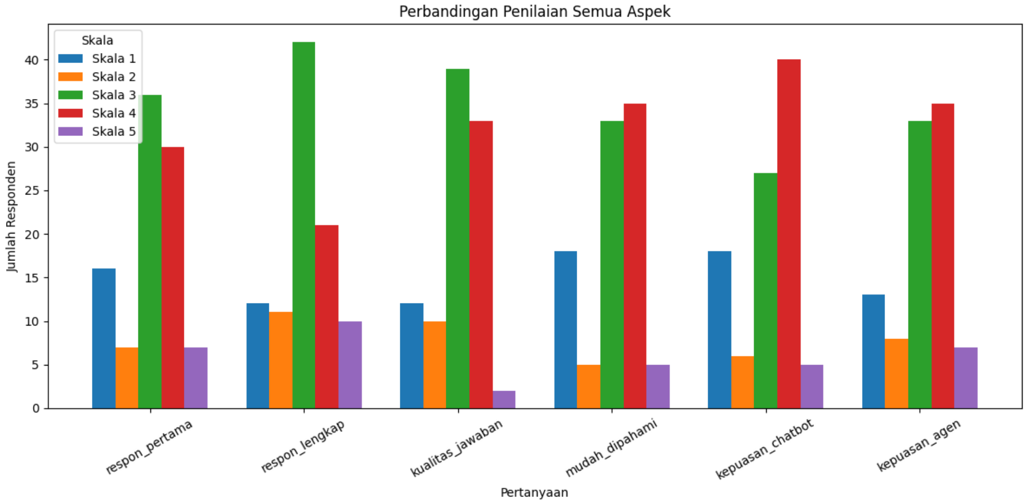
Alur komunikasi sistem chatbot ditunjukkan pada Gambar 4. Pengguna mengirimkan pesan melalui antarmuka frontend, yang kemudian diteruskan via HTTP POST pada backend Laravel. Backend meneruskan permintaan tersebut ke Chatbot API untuk diproses, dan respons dikembalikan ke frontend secara bertahap menggunakan mekanisme Server-Sent Events (SSE). Pendekatan streaming ini

ISSN 2502-3357 (*Online*)
F.F. Ashari dkk./TEKNOLOGI 16(1) 2026 31-42
ISSN 2503-0477 (*Print*)

memungkinkan token respons ditampilkan secara dinamis tanpa perlu penyegaran halaman, sehingga menghasilkan pengalaman interaksi yang lebih responsif bagi pengguna.

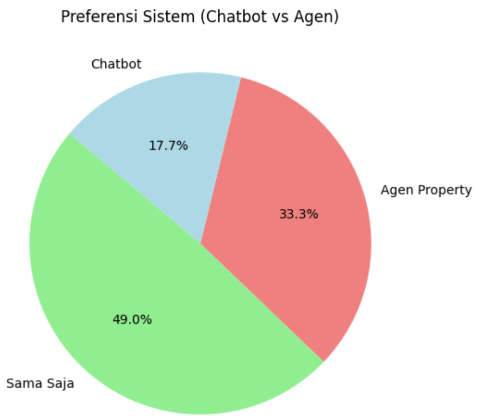
4.4 Uji Implementasi dan Evaluasi Pengguna

Pengujian implementasi model terhadap sistem dilakukan dengan membandingkan performa chatbot berbasis TinyLLaMA yang telah di-fine-tune dengan LoRA terhadap agen manusia melalui kuesioner pengguna. Responden mengajukan pertanyaan dengan salah satu dari tujuh topik yang telah ditentukan, kemudian mengisi kuesioner yang mengukur kecepatan respons, kualitas jawaban, kemudahan dipahami, dan tingkat kepuasan. Hasil perbandingan keseluruhan disajikan pada Gambar 6.



Gambar 6. Grafik Perbandingan Hasil Semua Kuesioner

Secara umum, penilaian responden pada seluruh aspek didominasi oleh Skala 3 (sedang/cukup) dan Skala 4 (lambat/tidak lengkap/sulit/tidak puas). Pada aspek kecepatan respons pertama, mayoritas responden menilai sedang (36 orang) dan lambat (30 orang), sementara pada kecepatan respons lengkap pola serupa terlihat dengan 41 orang menilai sedang dan 21 orang menilai lambat, mengindikasikan responsivitas sistem masih perlu dioptimalkan. Dari sisi kualitas jawaban, 39 orang menilai cukup lengkap dan 33 orang menilai tidak lengkap, sedangkan pada kemudahan dipahami 35 orang menilai sulit dan 33 orang menilai cukup mudah. Dominasi penilaian pada Skala 3 dan 4 ini sejalan dengan karakteristik output model yang diamati selama pengujian, di mana model sesekali menghasilkan jawaban dengan struktur kalimat yang kurang baku atau respons yang tidak sepenuhnya relevan dengan pertanyaan yang diajukan. Hanya pada sebagian kecil kasus model mampu memberikan jawaban yang cukup jelas, meskipun tetap masih terdapat kekurangan dalam kelengkapan informasi. Keterbatasan ini wajar mengingat kapasitas 1.1 miliar parameter dari model itu sendiri. Koto et al. (2023) menunjukkan model skala ini kesulitan memahami struktur bahasa Indonesia yang kompleks. Menariknya, tingkat kepuasan terhadap agen manusia menunjukkan distribusi yang relatif serupa dengan chatbot, di mana kepuasan chatbot didominasi tidak puas (40 orang) dan cukup puas (27 orang), sementara kepuasan agen manusia juga didominasi tidak puas (35 orang) dan cukup puas (33 orang), mengindikasikan bahwa performa keduanya berada pada level yang sebanding meski sama-sama belum optimal.



Gambar 7. Preferensi Sistem (Chatbot vs Agen)

Hasil preferensi sistem pada Gambar 7 menunjukkan 49.0% responden menilai chatbot dan agen manusia setara, 33.3% memilih agen properti, dan 17.7% memilih chatbot. Proporsi terbesar pada kategori "sama saja" ini tidak serta-merta mencerminkan bahwa chatbot sudah sangat baik, melainkan lebih mengindikasikan bahwa gap performa antara keduanya tidak cukup signifikan untuk dirasakan responden kemungkinan karena agen manusia yang menjadi pembanding juga belum memberikan layanan yang jauh lebih unggul dalam sesi pengujian tersebut. Temuan ini menunjukkan bahwa chatbot masih berada pada tahap awal yang memerlukan banyak penyempurnaan, terutama pada aspek konsistensi jawaban dan kejelasan bahasa, sehingga pendekatan hybrid antara chatbot dan agen manusia tetap direkomendasikan sebagai strategi layanan yang lebih realistis untuk kondisi saat ini.

5. Kesimpulan

Penelitian ini telah menunjukkan bahwa fine-tuning LoRA pada model TinyLLaMA-1.1B-Chat-v1.0 dengan dataset sintetis berbasis Template-Based Data Generation mampu menghasilkan chatbot konsultan properti berbahasa Indonesia yang fungsional, meskipun kualitas dialog yang dihasilkan masih memerlukan peningkatan lebih lanjut untuk mencapai standar layanan yang optimal. Eksperimen pada tiga skenario konfigurasi menunjukkan bahwa variasi nilai rank LoRA berpengaruh signifikan terhadap kualitas perilaku dialog yang dihasilkan. Skenario 3 dengan konfigurasi rank r=24 dan lora_alpha=48 menghasilkan performa terbaik dengan skor BLEU-4 sebesar 16.57, mengungguli Skenario 1 (15.27) dan Skenario 2 (13.39). Temuan ini mengonfirmasi bahwa peningkatan kapasitas representasi melalui rank yang lebih besar lebih efektif dalam meningkatkan kualitas dialog dibandingkan sekadar memperpanjang durasi pelatihan dengan regularisasi yang lebih ketat.

Hasil uji implementasi melalui kuesioner pengguna menunjukkan bahwa performa chatbot berada pada level yang sebanding dengan agen manusia, tercermin dari distribusi kepuasan yang relatif serupa antara keduanya. Sebanyak 49.0% responden menilai chatbot dan agen manusia setara dalam memberikan layanan, sementara preferensi terhadap agen manusia (33.3%) hanya sedikit lebih tinggi dibandingkan chatbot (17.7%). Proporsi terbesar pada kategori "sama saja" ini kemungkinan tidak semata-mata mencerminkan kualitas chatbot yang sudah baik, melainkan juga dipengaruhi oleh kondisi agen manusia yang dalam sesi pengujian belum memberikan layanan yang secara signifikan lebih unggul. Meskipun demikian, aspek kecepatan respons, kelengkapan jawaban, dan kemudahan dipahami masih perlu dioptimalkan, mengingat mayoritas penilaian responden masih berada pada kategori sedang hingga kurang memuaskan.

Berdasarkan temuan tersebut, pendekatan hybrid yang menggabungkan chatbot untuk menangani pertanyaan rutin dan agen manusia untuk pertanyaan yang lebih kompleks direkomendasikan sebagai strategi layanan yang lebih realistis untuk kondisi saat ini. Untuk penelitian selanjutnya, disarankan penggunaan teknik Retrieval-Augmented Generation (RAG) guna meningkatkan akurasi dan kedalaman jawaban terutama pada topik legal dan keuangan properti, perluasan dataset dengan cakupan skenario yang lebih beragam dan natural, penambahan metrik evaluasi seperti ROUGE-L atau BERTScore untuk melengkapi keterbatasan BLEU-4 dalam mengukur kualitas dialog, serta eksplorasi model dasar yang lebih besar untuk meningkatkan kemampuan generalisasi chatbot pada domain properti berbahasa Indonesia.

6. Kontribusi Penulis

F. F. Ashari: *Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, dan Writing – original draft*. **A. T. W. Almais:** *Investigation, Methodology, Software, dan Writing – original draft*. **F. R. Hariri:** *Data curation, Methodology, Visualization, dan Writing – original draft*. **F. F. Ashari:** *Software dan Visualization*. **A. T. W. Almais:** *Conceptualization, Funding acquisition, Supervision, Validation, dan Writing – review & editing*.

7. *Declaration of Competing Interest*

Penulis menyatakan tidak ada konflik kepentingan.

8. Referensi

Dettmers, T., & May, L. G. (n.d.). QL O RA : Efficient Finetuning of Quantized LLMs.
Ding, B., Qin, C., et al. (2023) – Data Augmentation using Large Language Models

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Faturahman, M. A., Studi, P., & Informasi, S. (2023). PERANCANGAN APLIKASI SISTEM INFORMASI AGEN REAL ESTATE BERBASIS WEBSITE DI PERUSAHAAN BROKER PERUMAHAN PT BERKAT ANUGERAH PRIMA KOTA BANDUNG. 1–15. https://doi.org/10.32897/infotronik.202x.6.2.1566

Fu, X., Laskar, T. R., Khasanova, E., Chen, C., & Tn, S. B. (2024). Tiny Titans : Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization ?

Herlawati, H., & Handayanto, R. T. (2026). Fine-Tuning Large Language Model (LLM) for Chatbot with Additional Data Sources. 13(225), 125–132. https://doi.org/10.33558/piksel.v13i1.10832

Huan, M., & Shun, J. (2025). Fine-Tuning Transformers Efficiently : A Survey on LoRA and Its Impact Fine-Tuning Transformers Efficiently : A Survey on LoRA and Its Impact. 0–17. https://doi.org/10.20944/preprints202502.1637.v1

Liu, J., Kong, Z., Dong, P., Yang, C., Shen, X., Zhao, P., Tang, H., Yuan, G., Niu, W., Zhang, W., Lin, X., Huang, D., & Wang, Y. (n.d.). RoRA : Efficient Fine-Tuning of LLM with Reliability Optimization for Rank Adaptation.

Maheshwari, G., Ivanov, D., & Haddad, K. El. (n.d.). Efficacy of Synthetic Data as a Benchmark.

Meyer, S., Singh, S., Tam, B., Ton, C., & Ren, A. (2024). A Comparison of LLM Fine-tuning Methods and Evaluation Metrics with Travel Chatbot Use Case. 1–28.

Pandya, K., & Mahavidyalaya, B. V. (2023). Automating Customer Service using LangChain. 28–31.

Paula, R. T., Neto, G. A., Romero, D., & Guerra, P. T. (n.d.). Evaluation of Synthetic Datasets Generation for Intent Classification Tasks in Portuguese.

Pratap, S., Richard, A., Kumar, D., & Malhotra, G. (2025). The fine art of fine-tuning : A structured review of advanced LLM fine-tuning techniques. Natural Language Processing Journal, 11(October 2024), 100144. https://doi.org/10.1016/j.nlp.2025.100144

Rahmaddion, A., & Arribe, E. (2024). PERANCANGAN SISTEM INFORMASI PENJUALAN RUMAH BERBASIS WEB PADA PT . AGUNG SELARAS GROUP PEKANBARU. 2–7.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., & Khashabi, D. (2023). : Aligning Language Models with Self-Generated Instructions. Lm.

Wu, J., Zhu, Y., Shen, L., & Lu, X. (2024). GEB-1.3B: Open Lightweight Large Language Model.

Yang, A., Liu, K., Liu, J., Lyu, Y., & Li, S. (2018). Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task. 98–104.

Yu, H. Q., & Mcguinness, S. (2024). An Experimental Study of Integrating Fine-tuned LLMs and Prompts for Enhancing Mental Health Support Chatbot System.

Zhang, P., Wang, T., & Lu, W. (n.d.). TinyLlama : An Open-Source Small Language Model. 1–10.

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv:2106.09685. https://arxiv.org/abs/2106.09685

Zhang, Y., Lau, R. Y. K., Xu, J. D., Rao, Y., & Li, Y. (2024). Business chatbots with deep learning technologies : state - of - the - art , taxonomies , and future research directions. In Artificial Intelligence Review (Vol. 57, Issue 5). Springer Netherlands. https://doi.org/10.1007/s10462-024-10744-z

Koto, F., Aisyah, N., Li, H., & Baldwin, T. (2023). Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. arXiv. https://doi.org/10.48550/arXiv.2310.04928