

**OPTIMALISASI CHATBOT BERBASIS TINYLLAMA DENGAN
FINE - TUNING LORA PADA SISTEM LAYANAN
AGENSI PROPERTY**

SKRIPSI

**Oleh:
FACHRIZAL FAZZA ASHARI
NIM. 220605110077**



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

**OPTIMALISASI CHATBOT BERBASIS TINYLLAMA DENGAN
FINE - TUNING LORA PADA SISTEM LAYANAN
AGensi PROPERTY**

SKRIPSI

Diajukan Kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
FACHRIZAL FAZZA ASHARI
NIM. 220605110077

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

HALAMAN PERSETUJUAN

OPTIMALISASI CHATBOT BERBASIS TINYLLAMA DENGAN FINE - TUNING LORA PADA SISTEM LAYANAN AGensi PROPERTY

SKRIPSI

Oleh :
FACHRIZAL FAZZA ASHARI
NIM. 220605110077

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 10 Juni 2026

Pembimbing I,



Dr. Agung Teguh Wibowo Almais, M.T.
NIP. 19860301 202321 1 016

Pembimbing II,



Fajar Rohman Hariri, M.Kom.
NIP. 19890515 201801 1 001

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M.Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

OPTIMALISASI CHATBOT BERBASIS TINYLLAMA DENGAN FINE - TUNING LORA PADA SISTEM LAYANAN AGENSI PROPERTY

SKRIPSI

Oleh :
FACHRIZAL FAZZA ASHARI
NIM. 220605110077

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 17 Juni 2026

Susunan Dewan Penguji

Ketua Penguji	: <u>Dr. Zainal Abidin, M.Kom</u> NIP. 19760613 200501 1 004
Anggota Penguji I	: <u>Dr. Totok Chamidy, M.Kom</u> NIP. 19691222 200604 1 001
Anggota Penguji II	: <u>Dr. Agung Teguh Wibowo A., M.T</u> NIP. 19860301 202321 1 016
Anggota Penguji III	: <u>Fajar Rohman Hariri, M.Kom</u> NIP. 19890515 201801 1 001

()
()
()
()

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M.Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Fachrizal Fazza Ashari
NIM : 220605110077
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Optimalisasi Chatbot Berbasis Tinyllama Dengan
Fine - Tuning Lora Pada Sistem Layanan Agensi
Property

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 26 Juni 2026

Yang membuat pernyataan,



Fachrizal Fazza Ashari
NIM. 220605110077

MOTTO

"Ketika kamu memilih kebebasan, kamu juga memilih tanggung jawab."

-Richie Norton-

HALAMAN PERSEMBAHAN

Alhamdulillah rabbil 'alamin puji syukur kepada Allah SWT karena berkat rahmat dan petunjuk-Nya penulis dapat menyelesaikan skripsi ini. Shalawat serta salam selalu tercurahkan kepada Rasulullah SAW yang telah membawa kita dari zaman jahiliyah menuju addinul Islam.

Penulis mempersembahkan karya ini kepada diri orang tua, diri sendiri, kakak, para dosen, teman-teman, serta orang-orang yang telah membantu, mendoakan, serta menyemangati penulis dalam menuntaskan skripsi ini.

KATA PENGANTAR

Assalamu 'alaikum warahmatullahi wabarakatuh

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala limpahan rahmat, taufik, dan hidayah-Nya, sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul “Optimalisasi Chatbot Berbasis Tinyllama dengan Fine - Tuning Lora pada Sistem Layanan Agensi Property” dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penyusunan skripsi ini tentu bukanlah proses yang mudah. Berbagai tantangan dihadapi mulai dari proses pencarian topik penelitian, pengumpulan referensi, perancangan sistem, implementasi model, hingga proses pengujian dan evaluasi hasil penelitian. Namun, berkat bimbingan, dukungan, doa, serta semangat dari berbagai pihak, penulis dapat menyelesaikan seluruh proses penyusunan skripsi ini dengan penuh kesungguhan dan ketekunan.

Penulis menyadari bahwa tersusunnya skripsi ini tidak terlepas dari bantuan, arahan, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. H. Agus Mulyono, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

3. Supriyono, M.Kom, selaku Ketua Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Agung Teguh Wibowo Almais, M.Kom, selaku Pembimbing I dan Fajar Rohman Hariri, M.Kom, selaku Pembimbing II, yang dengan penuh kesabaran, perhatian, serta dedikasi telah membimbing penulis dalam setiap tahapan penyusunan skripsi ini. Segala arahan, masukan, kritik, dan ilmu yang diberikan menjadi bagian penting dalam proses penyelesaian penelitian ini.
5. Dr. Zainal Abidin, M.Kom, selaku Ketua Penguji dan Dr. Totok Chamidy, M.Kom, selaku Anggota Penguji I, yang telah memberikan kritik, saran, serta masukan konstruktif demi menyempurnakan kualitas penelitian dan penulisan skripsi penulis.
6. Seluruh dosen dan staf Program Studi Teknik Informatika, yang telah memberikan ilmu pengetahuan, pelayanan akademik, serta pengalaman berharga selama masa perkuliahan.

7. Kedua orang tua dan keluarga tercinta, atas doa, kasih sayang, dukungan moral maupun material, serta semangat yang tidak pernah putus dalam setiap perjalanan pendidikan penulis. Pengorbanan dan dukungan mereka menjadi kekuatan terbesar bagi penulis untuk terus melangkah hingga menyelesaikan pendidikan ini.
8. Muhammad Faqih, Wafiy Anwarul Hikam, Ridho Aulia Rahman, dan Maulana Haekal Akbar yang senantiasa menjadi rekan berdiskusi dan bertukar pemikiran selama masa perkuliahan hingga proses penyelesaian skripsi ini. Terima kasih atas dukungan, semangat, motivasi, serta berbagai pandangan dan pengalaman yang telah dibagikan, baik dalam bidang akademik maupun kehidupan sehari-hari, sehingga memberikan banyak pelajaran dan dorongan bagi penulis dalam menyelesaikan studi ini.
9. Seseorang yang tidak dapat penulis sebutkan namanya, yang telah memberikan dukungan, perhatian, dan semangat kepada penulis, terutama pada masa-masa yang paling sulit selama proses perkuliahan dan penyusunan skripsi ini. Terima kasih karena telah menjadi tempat berbagi cerita, keluh kesah, serta memberikan motivasi yang membantu penulis untuk tetap kuat, bangkit, dan terus melangkah hingga mampu menyelesaikan seluruh proses pendidikan ini. Semoga segala kebaikan yang telah diberikan mendapatkan balasan terbaik dari Allah SWT.
10. Seluruh pihak yang telah terlibat, baik secara langsung maupun tidak langsung dari awal perkuliahan hingga akhir penulisan skripsi ini.

Penulis berharap semoga skripsi ini dapat memberikan manfaat di masa yang akan datang. Penulis juga menyadari bahwa karya ini masih memiliki kekurangan. Oleh karena itu, penulis sangat terbuka terhadap segala bentuk masukan dan saran yang membangun untuk pengembangan lebih lanjut.

Wassalamu 'alaikum warahmatullahi wabarakatuh.

Malang, 16 April 2026

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN	iv
PERNYATAAN KEASLIAN TULISAN	v
MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	xii
DAFTAR GAMBAR	xiv
DAFTAR TABEL	xv
ABSTRAK	xvi
ABSTRACT	xvii
مستخلص البحث.....	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah	7
1.3 Batasan Penelitian	7
1.4 Tujuan Penelitian	8
1.5 Manfaat Penelitian.....	8
BAB II STUDI PUSTAKA	6
2.1 Penelitian Terkait.....	6
2.2 Tinjauan Topik Tanya Jawab dalam Lingkup Agensi Property	11
2.3 Tinjauan Template-Based Data Generation.....	14
2.4 Tinjauan Format Dataset.....	17
2.5 Arsitektur Transformer	19
2.6 Tinjauan Model TinyLLaMA	21
2.7 Tinjauan Metode Low-Rank Adaptation	23
2.8 Metode Evaluasi Model dengan BLEU.....	25
2.9 Metode Uji Implementasi Model	28
BAB III DESAIN DAN IMPLEMENTASI.....	30
3.1 Desain Penelitian	30
3.2 Data Preparation	31
3.2.1 <i>Template-Based Data Generation</i>	32
3.2.2 Evaluasi Kualitas Dataset dengan Quesioner	40
3.3 Desain Sistem	41
3.3.1 <i>Dataset</i>	42
3.3.2 <i>Low Rank Adaptation</i>	43
3.3.3 Model TinyLlama-1.1B-Chat-v1.0	46
3.3.4 Cross Entropy Loss	50
3.3.5 Backpropagation	50
3.3.6 LoRA Parameter	52
3.4 Skenario Uji coba	53
3.4.1 Konfigurasi Pelatihan Model	54

3.4.2 Uji Hasil Pelatihan Model.....	56
3.4.3 Pemilihan Model Terbaik.....	56
3.4.4 Uji Implementasi Model	57
3.5 Evaluasi BLEU	58
3.6 Evaluasi Pengguna	60
3.7 Implementasi Model Chatbot	61
3.8 Contoh Perhitungan Manual Fine-Tuning LoRA.....	64
3.8.1 Representasi Input (Embedding)	65
3.8.2 Bobot Pretrained Query (Dibekukan)	65
3.8.3 Parameter LoRA	66
3.8.4 Perhitungan Update Low-Rank (BA)	66
3.8.5 Pembentukan Bobot Efektif.....	67
3.8.6 Forward Pass: Perhitungan Query	67
3.8.7 Perhitungan Logits dan Softmax.....	68
3.8.8 Perhitungan Cross-Entropy Loss	69
3.8.9 Perhitungan Gradien	69
3.8.10 Update Parameter LoRA.....	70
BAB IV HASIL DAN PEMBAHASAN.....	71
4.1 Gambaran Umum Penelitian	71
4.2 Dataset Pelatihan	72
4.2.1 Template-Based Data Generation	73
4.2.2 Ringkasan Akhir Dataset	82
4.3 Pelatihan Model.....	83
4.3.1 Konfigurasi Pelatihan Model	84
4.3.2 Hasil Proses Pelatihan Model	87
4.3.3 Evaluasi Kualitas Respons Model	92
4.3.4 Analisis Model Terbaik	94
4.4 Implementasi Model Pada Sistem	96
4.5 Uji Implementasi Model.....	99
4.5.1 Pembahasan Hasil Uji Implementasi Model.....	100
4.5.2 Analisis Apakah Model Sudah Optimal.....	104
4.6 Integrasi Islam	106
BAB V KESIMPULAN DAN SARAN	110
5.1 Kesimpulan.....	110
5.2 Saran.....	111
5.2.1 Saran Pengembangan Sistem	111
5.2.2 Saran Penelitian Selanjutnya	112
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 2. 1 Contoh Format Dataset	18
Gambar 3. 1 Desain Penelitian	30
Gambar 3. 2 Alur Template-Based Data Generation	32
Gambar 3. 3 Desain Sistem	42
Gambar 3. 4 Self Attention.....	47
Gambar 3. 5 Alur Menjalankan Chatbot	62
Gambar 3. 6 Flowchart Alur Komunikasi Chatbot pada Sistem	62
Gambar 4. 1 Grafik Training Loss vs Step Skenario 1.....	88
Gambar 4. 2 Grafik Training Loss vs Step Skenario 2	89
Gambar 4. 3 Grafik Training Loss vs Step Skenario 3	91
Gambar 4. 4 Sequence Diagram Alur Komunikasi antara Pengguna dengan Chatbot.....	96
Gambar 4. 5 Landing Page Website	98
Gambar 4. 6 Tampilan UI Chat antara User dengan Chatbot.....	99

DAFTAR TABEL

Tabel 2. 1 Tabel Penelitian Terkait	7
Tabel 2. 2 Intent Dataset.....	12
Tabel 3. 1 List Intent Dataset	33
Tabel 3. 2 Format Dataset.....	34
Tabel 3. 3 Prompt Generator Dataset	36
Tabel 3. 4 Hasil Generate Dataset	39
Tabel 3. 5 Skenario Uji.....	54
Tabel 3. 6 Contoh Token Perhitungan.....	59
Tabel 3. 7 Contoh Perhitungan BLEU-4	59
Tabel 3. 8 Contoh Quesioner.....	60
Tabel 3. 9 Permisalan Nilai Logits	68
Tabel 4. 1 Data Hasil Generate.....	78
Tabel 4. 2 Penilaian Kualitas dan Relevansi Dataset Chatbot Properti	79
Tabel 4. 3 Tabel Jumlah Intent pada Dataset	82
Tabel 4. 4 Format Dataset.....	83
Tabel 4. 5 Konfigurasi Pelatihan Skenario 1	85
Tabel 4. 6 Konfigurasi Pelatihan Skenario 2	86
Tabel 4. 7 Konfigurasi Pelatihan Skenario 3	86
Tabel 4. 8 Evaluasi Metriks Skenario 1	92
Tabel 4. 9 Evaluasi Metriks Skenario 2	93
Tabel 4. 10 Evaluasi Metriks Skenario 3	94
Tabel 4. 11 Data Hasil Kuesioner Skala Likert per Responden	100
Tabel 4. 12 Statistik Deskriptif Empat Variabel Pengukuran.....	102
Tabel 4. 13 Pedoman Interval Skor Skala Likert	102

ABSTRAK

Ashari, Fachrizal Fazza. 2026. **Optimalisasi Chatbot Berbasis TinyLLaMA dengan Fine-Tuning LoRA pada Sistem Layanan Agensi Property**. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Agung Teguh Wibowo Almais, M.T (II) Fajar Rohman Hariri, M.Kom.

Kata Kunci: *chatbot, TinyLLaMA, fine-tuning LoRA, agensi property.*

Penelitian ini berangkat dari dua permasalahan utama pada layanan agensi properti. Pertama, sistem konvensional yang mengandalkan interaksi manual dan agen manusia memiliki keterbatasan dalam kecepatan respons dan ketersediaan layanan. Keterbatasan model LLM dalam memberikan respons yang sesuai domain, personal, dan efisien dalam konteks properti berbahasa Indonesia, sehingga menyebabkan kesenjangan layanan antara kebutuhan pelanggan modern dengan sistem konvensional. Penelitian bertujuan mengadaptasi model TinyLLaMA-1.1B-Chat-v1.0 menjadi chatbot konsultan properti berbahasa Indonesia melalui fine-tuning LoRA, serta menganalisis pengaruh variasi nilai rank terhadap kualitas respons. Dataset sintetis sebanyak 3024 pasang instruction-output dihasilkan dengan metode Template-Based Data Generation mencakup tujuh intent, yang telah melalui validasi oleh lima responden dari bidang properti. Proses fine-tuning dilakukan pada tiga skenario nilai LoRA Rank ($r=8$, $r=16$, $r=32$) dengan LoRA Alpha ($\alpha=2r$) masing-masing 16, 32, dan 64, serta parameter lain dikontrol identik. Hasil evaluasi dengan BLEU-4 menunjukkan skor tertinggi 14.54 dicapai Skenario 3 ($r=32$), hanya berbeda 0.05 poin dari Skenario 2 ($r=16$) yang memperoleh 14.49, sementara Skenario 1 ($r=8$) memperoleh 12.76. Uji implementasi terhadap 45 responden menunjukkan keempat variabel pengukuran berada pada kategori Cukup dengan rata-rata 2.80–2.87, dengan kepuasan terhadap chatbot dan agen manusia setara (2.80). Penelitian ini menyimpulkan bahwa fine-tuning LoRA mampu mengadaptasi TinyLLaMA ke domain properti Indonesia, namun belum mencapai hasil optimal mengingat skor BLEU-4 yang rendah dan tingkat kepuasan yang hanya kategori Cukup. Peningkatan rank di atas $r=16$ juga tidak memberikan perbaikan substansial terhadap kesesuaian tekstual.

ABSTRACT

Ashari, Fachrizal Fazza. 2026. **Optimalisasi Chatbot Berbasis TinyLLaMA dengan Fine-Tuning LoRA pada Sistem Layanan Agensi Property**. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Agung Teguh Wibowo Almais, M.T (II) Fajar Rohman Hariri, M.Kom.

Keywords: *chatbot, TinyLLaMA, LoRA fine-tuning, property agency.*

This research stems from two main problems in property agency services. First, conventional systems that rely on manual interaction and human agents have limitations in response speed and service availability. Second, large language models (LLMs) face limitations in delivering responses that are domain-relevant, personalized, and efficient within the context of the Indonesian language property sector, thereby creating a service gap between modern customer needs and conventional systems. This study aims to adapt the TinyLLaMA-1.1B-Chat-v1.0 model into an Indonesian-language property consultant chatbot through LoRA fine-tuning, and to analyze the effect of rank variations on response quality. A synthetic dataset of 3024 instruction-output pairs was generated using the Template-Based Data Generation method covering seven intents, which was validated by five respondents from the property sector. Fine-tuning was conducted across three scenarios with LoRA Rank values ($r=8$, $r=16$, $r=32$) and LoRA Alpha ($\alpha=2r$) of 16, 32, and 64 respectively, while other parameters were controlled identically. BLEU-4 evaluation results showed the highest score of 14.54 achieved by Scenario 3 ($r=32$), only 0.05 points higher than Scenario 2 ($r=16$) which scored 14.49, while Scenario 1 ($r=8$) scored 12.76. The implementation test on 45 respondents showed that all four measured variables fell into the "Sufficient" category, with averages ranging from 2.80 to 2.87, while satisfaction with the chatbot and human agents was equal (2.80). This study concludes that LoRA fine-tuning is able to adapt TinyLLaMA to the Indonesian property domain, although it has not yet achieved optimal results given the low-to-moderate BLEU-4 scores and satisfaction levels that only reached the "Sufficient" category. Increasing the rank beyond $r=16$ also did not yield a substantial improvement in textual alignment.

مستخلص البحث

أشهري، فخرينال فزا. 2026. تحسين أداء دردشة آلية على نموذج TinyLLaMA من خلال تكييف الدقيق بتقنية LoRA في نظام خدمات الوكالات العقارية في بحث جامعي. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: الدكتور أغونغ تيغوه ويوويو ألبايس، الماجستير. المشرف الثاني: فجر رحمن حريري، الماجستير.

الكلمات الأساسية: دردشة آلية، TinyLLaMA، تكييف الدقيق بتقنية LoRA، الوكالات العقارية.

ينطلق هذا البحث من مشكلتين رئيسيتين في خدمات الوكالات العقارية. تتمثل الأولى أن الأنظمة التقليدية التي تعتمد على التفاعل اليدوي والوكلاء البشريين تعاني من محدودية سرعة الاستجابة وتوافر الخدمة. أما الثانية فتتمثل في محدودية نماذج اللغات الكبيرة (LLMs) في تقديم استجابات ملائمة للمجال العقاري، وشخصية، وفعالة في سياق العقارات باللغة الإندونيسية، مما يؤدي إلى وجود فجوة خدمية بين احتياجات العملاء المعاصرين والأنظمة التقليدية. ويهدف هذا البحث إلى تكييف نموذج TinyLLaMA-1.1B-Chat-v1.0 لتصبح دردشة آلية استشاريًا في مجال العقارات باللغة الإندونيسية من خلال تكييف الدقيق بتقنية LoRA، فضلًا عن تحليل أثر تنوع قيم الرتبة (Rank) في جودة الاستجابات. تكوّن مجموعة البيانات الاصطناعية من 3024 زوجًا من تعليمات المخرجات، وقد أنشئت باستخدام أسلوب توليد البيانات المعتمد على النماذج، وشملت سبعة أنواع من المقاصد (Intents)، كما خضعت لمراجعة من قبل خمسة مشاركين من الوكالات العقارية. وأجريت عملية تكييف الدقيق في ثلاثة سيناريوهات مختلفة لقيم رتبة LoRA، وهي (r=8) و (r=16) و (r=32)، مع ضبط قيمة LoRA Alpha ($\alpha=2r$) لتكون 16 و 32 و 64 على التوالي، مع تثبيت بقية المعلمات. وأظهرت نتائج التقييم باستخدام مقياس BLEU-4 أن أفضل نتيجة بلغت 14.54 قد تحققت في السيناريو الثالث (r=32)، بفارق قدره 0.05 نقطة فقط عن السيناريو الثاني (r=16) الذي حقق نتيجة قدرها 14.49، في حين حقق السيناريو الأول (r=8) نتيجة بلغت 12.76. كما أظهر تجربة التطبيق الذي شمل 45 مشاركًا أن المتغيرات الأربعة المقاسة جميعها وقعت ضمن فئة مقبول، بمتوسط تراوح بين 2.80 و 2.87، مع تساوي مستوى الرضا بين دردشة آلية والوكيل البشري بمتوسط قدره (2.80). ويخلص البحث إلى أن تكييف الدقيق بتقنية LoRA قادر على نموذج TinyLLaMA مع الوكالات العقارية في إندونيسيا، إلا أن النتائج ما زالت دون المستوى المأمول، نظرًا لانخفاض قيمة BLEU-4 واقتصار مستوى رضا المستخدمين على فئة مقبول. كما أن زيادة قيمة الرتبة إلى ما فوق (r=16) لم تؤدّ إلى تحسن جوهري في مستوى التوافق النصي.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi telah membawa perubahan besar dalam berbagai aspek kehidupan manusia, termasuk dalam dunia bisnis (Herlawati & Handayanto, 2026). Transformasi digital tidak hanya mengubah cara kita berkomunikasi dan mengakses informasi, tetapi juga merevolusi model bisnis tradisional menjadi lebih adaptif terhadap tuntutan pasar yang dinamis (Yang et al., 2025). Di era ini, setiap sektor ekonomi dipaksa untuk menyesuaikan diri dengan sistem yang lebih cepat, efisien, dan mudah diakses melalui perangkat digital.

Industri properti, sebagai salah satu sektor penting dalam perekonomian global dan nasional, tidak terkecuali dari tuntutan ini (Rahmaddion & Arribe, 2024). Sektor ini melibatkan transaksi bernilai tinggi yang memerlukan kecepatan respons informasi dan jangkauan luas untuk menarik calon pembeli. Namun, model pemasaran tradisional yang masih mengandalkan interaksi tatap muka, media cetak seperti brosur dan koran, serta pameran fisik semakin dianggap kurang relevan di era digital (Faturahman et al., 2023).

Keterbatasan metode ini terlihat jelas dalam kemampuannya yang terbatas untuk menjangkau konsumen secara luas, terutama di tengah mobilitas tinggi masyarakat modern, serta keterlambatan dalam penyajian informasi yang sering kali membuat peluang bisnis hilang karena persaingan yang ketat. Fenomena ini juga terjadi secara nyata di Indonesia, di mana industri properti tumbuh pesat

seiring dengan urbanisasi dan peningkatan kebutuhan hunian (Faturahman et al., 2023).

Kondisi serupa juga dialami oleh beberapa perusahaan properti yang masih menggunakan proses penjualan manual berbasis kertas dan catatan manual (Rahmaddion & Arribe, 2024). Keterbatasan ini menciptakan kesenjangan antara kebutuhan pasar akan akses cepat, transparan, dan personal dengan sistem operasional yang konvensional. Kondisi ini menegaskan perlunya pengembangan chatbot dengan kemampuan pada domain khusus yang mampu meningkatkan efisiensi komunikasi dan layanan pelanggan dalam industri properti.

Kesenjangan ini tidak hanya menghambat pertumbuhan bisnis, tetapi juga menurunkan kepuasan pelanggan, yang pada akhirnya berdampak pada daya saing perusahaan di pasar properti yang kompetitif. Upaya mengatasi kesenjangan tersebut telah dilakukan melalui berbagai inovasi, salah satunya adalah pengembangan sistem informasi berbasis web. Penelitian terdahulu menunjukkan bahwa sistem ini efektif dalam mengelola data properti secara terintegrasi, mempermudah proses pemesanan online, serta meminimalkan kesalahan pencatatan melalui otomatisasi (Rahmaddion & Arribe, 2024).

Misalnya, sistem berbasis web memungkinkan agen properti untuk memperbarui informasi stok unit, harga, dan spesifikasi secara cepat dan mudah, sehingga calon pelanggan dapat mengakses informasi property tanpa harus menghubungi agen secara langsung. Selain itu, fitur seperti pencarian filter berdasarkan lokasi, harga, atau tipe properti membuat proses eksplorasi lebih efisien. Namun, meskipun sistem ini telah membawa kemajuan, tantangan

berikutnya muncul dalam bentuk kebutuhan akan interaksi pelanggan secara real-time yang lebih personal dan responsif.

Pelanggan modern tidak hanya menginginkan informasi statis, tetapi juga konsultasi dinamis cepat yang bisa menjawab pertanyaan spesifik, seperti rekomendasi properti sesuai dengan jumlah uang yang dimiliki pelanggan. Selain dialog real-time, ringkasan percakapan otomatis terbukti membantu menjaga konteks layanan dan mempercepat respons pada skenario e-commerce (Han et al., 2020). Di sinilah teknologi kecerdasan buatan (AI), khususnya chatbot berbasis Large Language Models (LLMs), menawarkan potensi besar untuk mengisi celah tersebut. LLM, seperti yang digunakan dalam model seperti GPT atau LLaMA, mampu menghasilkan teks alami layaknya manusia, sehingga dapat meningkatkan kualitas komunikasi antara agen dan pelanggan dengan cara yang lebih engaging dan efisien (Herlawati & Handayanto, 2026).

Chatbot semacam ini bisa menangani pertanyaan rutin, memberikan saran personal. Meskipun demikian, penerapan LLM dalam domain spesifik seperti industri properti masih menghadapi keterbatasan yang tidak bisa diabaikan. Hal ini disebabkan karena model LLM pada dasarnya dilatih pada data umum yang bersifat generik, seperti teks dari internet atau buku, sehingga dibutuhkan penyesuaian perilaku dialog agar chatbot berperan sebagai agen properti yang komunikatif dan informatif (Yu & McGuinness, 2024).

Misalnya, LLM standar mungkin gagal memberikan jawaban akurat tentang rekomendasi atau informasi properti berdasarkan kondisi ekonomi pelanggan. Untuk menjawab tantangan ini, pendekatan fine-tuning dengan Low-Rank

Adaptation atau disingkat menjadi LoRA menjadi solusi yang efisien dan inovatif. LoRA memungkinkan penyesuaian LLM agar relevan dengan konteks tertentu tanpa membutuhkan sumber daya komputasi yang berlebihan, karena hanya melatih matriks low-rank tambahan daripada seluruh parameter model (Huan & Shun, 2025). Sejumlah studi melaporkan bahwa LoRA dapat menurunkan kebutuhan komputasi secara signifikan sambil menjaga kinerja model, sehingga relevan untuk skenario dengan keterbatasan anggaran.

Dalam penelitian ini, TinyLLaMA dipilih sebagai model dasar karena sifatnya yang ringan yaitu hanya sekitar 1.1B parameter, sehingga sesuai untuk implementasi pada perangkat keras terbatas. TinyLLaMA juga menunjukkan performa baik pada berbagai downstream tasks, dan dapat dioptimalkan melalui fine-tuning pada data domain-spesifik seperti deskripsi properti serta dialog pelanggan (Zhang et al., 2024). Penyesuaian perilaku dilakukan dengan pendekatan parameter-efisien seperti LoRA, yang efektif mengadaptasi LLM tanpa melatih seluruh parameter dan tanpa menambah latensi inferensi (Huan & Shun, 2025).

Upaya optimalisasi semacam ini, yaitu meningkatkan kualitas keluaran model agar lebih sesuai dengan kebutuhan pengguna, pada dasarnya sejalan dengan prinsip dalam Islam yang mendorong manusia untuk senantiasa memilih dan mengikuti hal yang terbaik dalam setiap upaya perbaikan. Islam mendorong manusia untuk menggunakan akal dan ilmu pengetahuan dalam menghasilkan solusi yang memberikan manfaat serta kualitas yang lebih baik bagi kehidupan. Prinsip tersebut tercermin dalam QS. Az-Zumar ayat 18:

الَّذِينَ يَسْتَمِعُونَ الْقَوْلَ فَيَتَّبِعُونَ أَحْسَنَهُ ۗ أُولَٰئِكَ الَّذِينَ هَدَاهُمُ اللَّهُ وَأُولَٰئِكَ هُمْ أُولُوا الْأَلْبَابِ (١٨)

“(Yaitu) mereka yang mendengarkan perkataan lalu mengikuti apa yang paling baik di antaranya. Mereka itulah orang-orang yang telah diberi petunjuk oleh Allah dan mereka itulah ululalbab (orang-orang yang mempunyai akal sehat).” (QS. Az-Zumar: 18)

Ayat ini menjelaskan tentang orang-orang yang mendengarkan berbagai perkataan, kemudian mengikuti yang paling baik di antaranya karena wahyu Allah merupakan petunjuk terbaik. Mereka adalah orang-orang yang memperoleh petunjuk dari Allah dan memiliki akal yang sehat serta pemikiran yang jernih (Kementerian Agama Republik Indonesia, n.d.-d).

Nilai yang terkandung dalam ayat tersebut memiliki relevansi dengan pengembangan sistem kecerdasan buatan. Namun demikian, kecerdasan buatan tidak dapat disamakan dengan manusia karena tidak memiliki akal, kesadaran, maupun kemampuan memahami ilmu sebagaimana manusia. Kecerdasan buatan hanya memproses pola berdasarkan data dan parameter yang dirancang serta disesuaikan oleh manusia. Dalam konteks penelitian ini, prinsip “mengikuti yang paling baik” tercermin pada upaya penelitian ini untuk mengoptimalkan model TinyLLaMA melalui metode Fine-Tuning LoRA agar menghasilkan respons yang lebih sesuai dengan kebutuhan layanan agen properti. Proses fine-tuning dilakukan bukan untuk memberikan akal kepada sistem, melainkan untuk menyesuaikan perilaku model sehingga mampu memberikan informasi yang lebih relevan, komunikatif, dan bermanfaat bagi pengguna.

Selain mendorong perkembangan teknologi, Al-Qur'an juga menekankan pentingnya dimensi etis dalam pengelolaan dan penyebaran informasi. Prinsip kehati-hatian dan verifikasi ditegaskan dalam QS. Al-Hujurat ayat 6:

يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ لَدْغَمِينَ (٦)

“Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu.” (QS. Al-Hujurat: 6)

Ayat ini memberikan landasan etis yang sangat relevan dalam konteks perkembangan teknologi informasi dan kecerdasan buatan. Berdasarkan tafsir Kementerian Agama Republik Indonesia, QS. Al-Hujurat ayat 6 mengajarkan agar tidak tergesa-gesa menerima suatu berita, khususnya yang berasal dari sumber yang tidak kredibel, tanpa terlebih dahulu meneliti kebenarannya, agar terhindar dari kesalahan dan penyesalan akibat tindakan yang didasarkan pada informasi yang belum tervalidasi (Kementerian Agama Republik Indonesia, n.d.-a). Prinsip tabayyun tersebut memiliki relevansi dalam pengembangan chatbot berbasis Large Language Model (LLM), di mana setiap informasi yang dihasilkan perlu melalui proses validasi dan evaluasi untuk meminimalkan kesalahan, bias, maupun informasi yang menyesatkan. Dengan demikian, pengembangan teknologi tidak hanya dituntut unggul secara teknis, tetapi juga harus memperhatikan tanggung jawab moral dan sosial dalam penyampaian informasi kepada pengguna.

Penelitian ini memfokuskan pada pembentukan gaya bahasa chatbot agen properti berbahasa Indonesia menggunakan Fine-Tuning LoRA pada TinyLLaMA. Data dinamis ditangani melalui pembaruan dataset dan pelatihan ulang adapter LoRA dari awal pada model TinyLLaMA dasar (tanpa adapter sebelumnya) untuk setiap versi dataset. Fokus bukan prediksi harga, melainkan gaya tutur, kepatuhan format jawaban, dan kejelasan langkah tindak lanjut.

Berdasarkan uraian di atas, penelitian ini berfokus pada pengembangan chatbot berbasis TinyLLaMA yang di fine-tune dengan LoRA. Implementasi ini diharapkan dapat meningkatkan efisiensi layanan secara keseluruhan, menyediakan informasi properti secara cepat melalui antarmuka yang intuitif, serta menghadirkan interaksi yang lebih responsif antara agen dan pelanggan.

1.2 Identifikasi Masalah

1. Apakah penerapan LoRA pada TinyLLaMA mampu mengoptimalkan chatbot agen properti berbahasa Indonesia dalam hal kesesuaian jawaban terhadap data rujukan?
2. Bagaimana pengaruh konfigurasi LoRA dengan menggunakan rank pada target layer q_proj , v_proj terhadap kualitas perilaku dialog yang diukur menggunakan metrik BLEU-4?

1.3 Batasan Penelitian

1. Penelitian berfokus pada pembentukan perilaku dialog chatbot agen properti berbahasa Indonesia, tanpa mencakup prediksi harga.
2. Metode fine-tuning yang digunakan dibatasi pada teknik Low-Rank Adaptation, dengan target layer adaptasi hanya diterapkan pada Query layer dan Value layer, tanpa membandingkan dengan metode Parameter-Efficient Fine-Tuning lainnya seperti QLoRA, Prefix Tuning, atau Adapter Tuning.
3. Penelitian ini tidak mencakup mekanisme pembaruan pengetahuan secara real-time. Pembaruan informasi hanya dapat dilakukan melalui proses pelatihan ulang model secara offline menggunakan dataset yang diperbarui.

4. Dataset yang digunakan dalam penelitian ini bersifat sintetis dan terbatas pada domain agensi properti berbahasa Indonesia, sehingga kemampuan generalisasi model di luar domain tersebut tidak dijamin.
5. Evaluasi berfokus pada kualitas dan kecepatan chatbot saat merespons tanpa evaluasi biaya, numerik, atau metrik bisnis.
6. Aspek safety, toxicity dan kepatuhan hukum tidak dievaluasi mendalam dimana nantinya hal ini akan menjadi pekerjaan lanjutan.

1.4 Tujuan Penelitian

1. Penelitian ini bertujuan menilai efektivitas LoRA pada TinyLLaMA dalam meningkatkan kualitas jawaban chatbot agen properti berbahasa Indonesia, melalui ketepatan akurasi jawaban. Berdasarkan kuesioner pengguna, penelitian ini membandingkan performa chatbot dan agen manusia dalam memberikan informasi properti.
2. Menganalisis pengaruh konfigurasi LoRA dengan menggunakan rank pada target layer q_proj , v_proj terhadap kualitas perilaku dialog.

1.5 Manfaat Penelitian

1. Bagi akademisi dan peneliti di bidang informatika: Memberikan contoh implementasi fine-tuning LoRA pada model ringan seperti TinyLLaMA untuk domain spesifik, yang dapat menjadi referensi untuk penelitian serupa di sektor lain seperti e-commerce atau layanan kesehatan.

2. Bagi pengembangan ilmu secara luas: Berkontribusi pada kemajuan teknologi AI yang efisien untuk perangkat dengan sumber daya terbatas, mendorong adopsi AI di perusahaan kecil hingga ke menengah di Indonesia.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian ini yang berfokus pada optimalisasi chatbot berbasis TinyLLaMA dengan fine-tuning LoRA untuk sistem layanan agensi properti, peninjauan terhadap penelitian terdahulu dilakukan untuk memperoleh pemahaman yang komprehensif mengenai perkembangan teknologi, metode, dan pendekatan yang relevan. Kajian ini tidak hanya bertujuan untuk mengidentifikasi berbagai solusi yang telah dikembangkan oleh peneliti sebelumnya, tetapi juga untuk memahami kelebihan, keterbatasan, serta tantangan yang masih dihadapi dalam pengembangan chatbot berbasis Large Language Model (LLM). Dengan demikian penelitian terdahulu dapat menjadi landasan konseptual dan teknis dalam menentukan pendekatan yang paling sesuai untuk menjawab permasalahan yang diangkat pada penelitian ini.

Peninjauan literatur diarahkan pada beberapa seperti perkembangan metode parameter-efficient fine-tuning (PEFT), khususnya LoRA dan variannya, penerapan model LLM dalam tugas percakapan, serta implementasi chatbot pada domain spesifik. Melalui kajian tersebut, dapat diperoleh gambaran mengenai perkembangan penelitian, metode yang digunakan, serta keterbatasan yang masih perlu diteliti lebih lanjut. Temuan ini menjadi dasar dalam merumuskan posisi dan kontribusi penelitian yang dilakukan.

Tabel 2. 1 Tabel Penelitian Terkait

No	Referensi	Topik	Metode	Subjek
1	(Meyer et al., 2024)	Komparasi fine-tuning	Komparasi QLoRA, RAFT, dan RLHF	LLaMA 2 7B dan Mistral 7B pada dataset travel Reddit (~16.300 entri)
2	(Liu et al., 2025)	Peningkatan LoRA dengan rank-adaptive	RoRA (variasi LoRA)	LLaMA-7B/13B, LLaMA2-7B, LLaMA3-8B, Sheared-LLaMA pada dataset 8 tugas commonsense
3	(Pratap et al., 2025)	Tinjauan chatbotLLM	Literature review 6 kategori fine-tuning	LLaMA, Mistral, GPT-3, RoBERTa, DeBERTa pada GLUE, MMLU, HellaSwag, GSM8K
4	(Hu et al., 2021)	Parameter- efficient fine- tuning	LoRA	GPT-3 175B, RoBERTa, DeBERTa, GPT-2 pada GLUE, E2E, WikiSQL, MNLI, SAMSum
5	(Ye et al., 2024)	Multi-GPU LoRA	mLoRA	TinyLlama-1.1B, Llama-2-7B/13B pada dataset GSM8K dengan GPU RTX A6000/3090
6	(Zhang et al., 2024)	Model ringan untuk percakapan	Arsitektur Llama 2	TinyLlama 1.1B pada SlimPajama (627B) + StarCoder; eval commonsense, MMLU, BBH, HumanEval, DROP
7	(Dettmers et al., 2023)	Fine-tuning dengan kuantisasi 4-bit	QLoRA	LLaMA 7B-65B, RoBERTa-large, T5 pada GLUE, MMLU, Vicuna, OASST1
8	(Xia et al., 2024)	Structured pruning untuk percepat pre-training	Targeted Structured Pruning + Dynamic Batch Loading	LLaMA2-7B → Sheared-LLaMA 1.3B/2.7B pada dataset RedPajama
9	(Pandya & Holia, 2023)	Custom GPT chatbot dengan LangChain	LLM + Langchain	Flan-T5 XXL/Base/Small pada data website BVM Engineering College
10	(Basilico, 2024)	Chatbot domain akuntansi, hukum, dan perpajakan Italia	QLoRA	Mistral 7B pada dataset legal-fiskal Italia (25.681 sampel)

Pada penelitian Hu et al. (2021) memperkenalkan LoRA sebagai metode parameter-efficient yang menjadi fondasi bagi hampir seluruh metode PEFT yang berkembang setelahnya. Penelitian ini bertujuan mengatasi keterbatasan full fine-tuning pada LLM dengan cara membekukan bobot pre-trained dan menyisipkan matriks dekomposisi rank rendah pada lapisan seluruh lapisan. Pengujian dilakukan pada berbagai model seperti RoBERTa, DeBERTa, GPT-2, dan GPT-3 175B dengan

benchmark GLUE, E2E, WikiSQL, dan SAMSum. Hasilnya menunjukkan bahwa LoRA dapat menurunkan jumlah parameter trainable hingga sepuluh ribu kali lipat tanpa mengorbankan performa, bahkan tanpa menambah latensi inferensi karena bobot adapter dapat digabungkan kembali ke model dasar. Penelitian ini menjadi landasan teoritis utama bagi penelitian yang dilakukan, terutama dalam hal pemilihan lapisan target adaptasi dan penggunaan scaling factor.

Seiring berkembangnya kebutuhan akan efisiensi yang lebih tinggi, Dettmers et al. (2023) mengusulkan QLoRA, yaitu kombinasi antara kuantisasi 4-bit dan LoRA yang memungkinkan fine-tuning LLM pada satu unit GPU dengan memori terbatas. Melalui inovasi 4-bit NormalFloat, double quantization, dan paged optimizer, penelitian ini berhasil melatih model 65B parameter pada GPU tunggal 48GB tanpa penurunan performa dibandingkan fine-tuning 16-bit penuh. Pada tahun yang berdekatan, Xia et al. (2024) mengembangkan pendekatan berbeda melalui structured pruning pada Sheared-LLaMA, yang memangkas model LLaMA2-7B menjadi model berukuran 1.3B dan 2.7B parameter dengan biaya komputasi jauh lebih rendah dibandingkan pelatihan dari awal. Kedua penelitian ini menegaskan bahwa efisiensi sumber daya dapat dicapai baik melalui kuantisasi maupun pemangkasan struktural, sebagai alternatif maupun pelengkap dari pendekatan LoRA murni.

Pada arah pengembangan LoRA itu sendiri, Liu et al. (2025) mengusulkan RoRA sebagai perbaikan terhadap masalah penurunan performa LoRA ketika rank ditingkatkan. Melalui analisis matematis terhadap variansi gradien, RoRA mengganti scaling factor dari α/r menjadi $\alpha/\sqrt{r}$ sehingga performa model tetap stabil

bahkan meningkat seiring kenaikan rank. Pengujian pada delapan dataset commonsense reasoning menggunakan LLaMA-7B hingga Sheared-LLaMA menunjukkan bahwa RoRA secara konsisten mengungguli LoRA dan DoRA, khususnya pada rank tinggi. Penelitian ini memberikan wawasan penting mengenai pengaruh konfigurasi rank terhadap stabilitas pelatihan, yang menjadi salah satu variabel yang turut dipertimbangkan dalam penentuan skenario rank pada penelitian ini.

Dari sisi efisiensi komputasi pada skala yang lebih besar, Ye et al. (2024) mengembangkan mLoRA, sebuah sistem pipeline parallelism yang memungkinkan pelatihan beberapa adapter LoRA secara simultan pada lingkungan multi-GPU, dengan TinyLlama-1.1B sebagai salah satu model yang diuji. Hasil penelitian menunjukkan penurunan waktu pelatihan hingga 45% dibandingkan metode paralelisasi konvensional. Penelitian ini relevan terutama dalam menggambarkan karakteristik TinyLlama sebagai model yang ringan namun tetap kompatibel dengan skema adaptasi LoRA pada berbagai skala perangkat keras.

Berkaitan langsung dengan model dasar yang digunakan dalam penelitian ini, Zhang et al. (2024) memperkenalkan TinyLlama, sebuah model bahasa berukuran 1,1 miliar parameter yang dilatih dengan data hingga tiga triliun token menggunakan arsitektur yang mengikuti Llama 2. Penelitian ini membuktikan bahwa model berukuran kecil dapat mencapai performa kompetitif, bahkan mengungguli OPT-1.3B dan Pythia-1.4B pada berbagai downstream task, melalui strategi pelatihan multi-tahap. Karakteristik ringan dari TinyLlama yang tetap mempertahankan kapasitas bahasa yang baik menjadi alasan utama pemilihan

model ini sebagai basis pengembangan chatbot pada penelitian ini, khususnya untuk mendukung implementasi pada perangkat dengan keterbatasan sumber daya komputasi.

Pratap et al. (2025) menyajikan tinjauan sistematis terhadap berbagai teknik fine-tuning LLM, mengklasifikasikannya ke dalam enam kategori utama mulai dari perubahan metode pelatihan hingga metode berbasis aplikasi. Tinjauan ini mencakup evaluasi terhadap berbagai model seperti LLaMA, Mistral, dan GPT pada benchmark seperti GLUE, MMLU, dan GSM8K, serta menegaskan bahwa keberhasilan fine-tuning sangat bergantung pada kesesuaian metode dengan tugas dan domain, dan bahwa evaluasi berbasis manusia tetap diperlukan karena metrik otomatis seperti BLEU dan ROUGE belum mampu menangkap kualitas interaksi secara menyeluruh. Temuan ini memberikan landasan penting bagi penelitian ini dalam mempertimbangkan keterbatasan metrik BLEU-4 sebagai satu-satunya metrik kuantitatif, sekaligus menjadi dasar argumentasi mengenai pentingnya evaluasi pengguna sebagai pelengkap.

Sejalan dengan hal tersebut, Meyer et al. (2024) melakukan studi komparatif terhadap tiga metode fine-tuning, yaitu QLoRA, RAFT, dan RLHF, pada kasus penggunaan chatbot perjalanan berbasis LLaMA 2 dan Mistral. Penelitian ini menemukan bahwa metrik otomatis seperti BLEU dan ROUGE tidak selalu selaras dengan evaluasi manusia, sementara evaluasi berbasis GPT-4 menunjukkan korelasi terbaik. Temuan ini memperkuat argumen bahwa penilaian kuantitatif semata tidak cukup untuk menggambarkan kualitas chatbot secara utuh.

Dua penelitian terakhir memberikan gambaran konkret mengenai implementasi chatbot domain spesifik. Pandya dan Holia (2023) membangun chatbot customer service menggunakan pendekatan RAG dengan LangChain dan Flan-T5, tanpa melibatkan proses fine-tuning, melainkan mengandalkan retrieval berbasis embedding dari data institusi pendidikan. Implementasi dilakukan dengan melakukan web scraping terhadap data FAQ dan informasi institusi, kemudian membangun sistem retrieval menggunakan FAISS yang dipadukan dengan model Flan-T5 dalam berbagai ukuran. Hasil pengujian menunjukkan bahwa varian Flan-T5-XXL memberikan kualitas respons yang lebih baik dibandingkan varian Base dan Small, dengan kemampuan menjawab pertanyaan secara kontekstual serta mempertahankan konteks percakapan melalui riwayat chat.

Berbeda dengan penelitian Pandya dan Holia, Basilico (2024) mengembangkan chatbot untuk firma akuntansi di Italia melalui fine-tuning pada model Mistral 7B menggunakan QLoRA, dengan dataset gabungan antara data hasil scraping dan data sintetis yang dihasilkan menggunakan LLM. Hasilnya menunjukkan bahwa model hasil fine-tuning mengungguli model dasar pada domain legal-fiskal, dengan biaya pelatihan yang sangat rendah. Penelitian ini memiliki kemiripan struktural yang erat dengan penelitian yang penulis lakukan, baik dari segi penggunaan dataset domain spesifik maupun pendekatan fine-tuning berbasis adapter pada model berukuran kecil.

2.2 Tinjauan Topik Tanya Jawab dalam Lingkup Agensi Property

Pada bagian ini membahas berbagai topik/intent yang relevan dalam konteks keputusan pembelian properti. Pembelian properti merupakan keputusan

besar yang melibatkan berbagai faktor, seperti harga, lokasi, kualitas produk, dan aspek hukum. Dalam hal ini, proses pembelian properti melibatkan pertimbangan tambahan terkait dengan kesesuaian transaksi yang ada di Indonesia, yang menjadi daya tarik bagi para konsumen. Beberapa topik/intent yang akan dibahas dalam penelitian ini meliputi simulasi KPR, legalitas dokumen, biaya dan pajak, spesifikasi properti, proses transaksi, konsultasi keputusan, serta risiko dan disclaimer yang perlu dipahami oleh konsumen.

Tabel 2. 2 Intent Dataset

Topik/Intent	Landasan Penelitian	Deskripsi
kpr_simulasi	Simulasi KPR adalah alat penting yang memungkinkan konsumen untuk menghitung kemampuan finansial mereka dalam membayar cicilan rumah. Hal ini menjadi aspek yang sangat penting dalam keputusan pembelian rumah karena memberikan gambaran yang jelas mengenai kemampuan konsumen untuk mengatur pembayaran properti dalam jangka panjang (Akhsya et al., 2021).	Pertanyaan atau permintaan yang berkaitan dengan perhitungan kemampuan finansial konsumen dalam membayar cicilan rumah melalui KPR, mencakup estimasi angsuran bulanan, uang muka, tenor, dan suku bunga.
legal_dokumen	Legalitas dalam transaksi properti sangat penting untuk memastikan bahwa pembelian rumah sah dan tidak bermasalah di kemudian hari. Dokumen seperti sertifikat tanah, IMB, dan perjanjian jual beli harus diperiksa dengan cermat oleh konsumen untuk menghindari potensi masalah hukum di masa depan (Septiana et al., 2020).	Pertanyaan yang menyangkut keabsahan dan kelengkapan dokumen hukum dalam transaksi properti, seperti sertifikat tanah (SHM/HGB), IMB, AJB, serta aspek legalitas lainnya yang diperlukan untuk memastikan kepemilikan rumah bebas dari sengketa.
biaya_dan_pajak	Biaya dan pajak yang terkait dengan pembelian properti merupakan faktor utama dalam keputusan pembelian. Biaya tambahan seperti pajak properti, biaya notaris, dan biaya administrasi lainnya harus dipertimbangkan dengan seksama agar konsumen tidak terjebak dalam masalah finansial setelah pembelian (Septiana et al., 2020).	Pertanyaan terkait seluruh komponen biaya dan kewajiban pajak dalam proses pembelian properti, meliputi pajak bumi dan bangunan (PBB), BPHTB, biaya notaris, biaya balik nama, serta biaya administrasi lainnya yang muncul di luar harga pokok properti.

spesifikasi_properti	Spesifikasi properti meliputi ukuran, desain, kualitas bangunan, dan fasilitas tambahan seperti taman atau sistem pengelolaan limbah yang baik. Konsumen cenderung memilih properti yang memenuhi kebutuhan fungsional dan estetika mereka (Akhsya et al., 2021).	Pertanyaan mengenai detail dan karakteristik fisik sebuah properti, mencakup luas bangunan dan tanah, jumlah kamar, desain arsitektur, material bangunan, serta fasilitas penunjang seperti taman, carport, dan sistem utilitas.
proses_transaksi	Proses transaksi pembelian rumah melibatkan beberapa langkah penting, seperti penentuan harga, pengaturan pembayaran, dan penandatanganan dokumen legal. Konsumen harus memastikan bahwa transaksi dilakukan secara transparan dan sesuai dengan hukum yang berlaku untuk menghindari masalah di kemudian hari (Septiana et al., 2020).	Pertanyaan seputar alur dan tahapan dalam proses jual beli properti, mulai dari penawaran harga, pembayaran tanda jadi (bookingfee), penandatanganan Perjanjian Pengikatan Jual Beli (PPJB), hingga serah terima kunci secara resmi.
konsultasi_keputusan	Sebelum membeli properti, konsumen sering berkonsultasi dengan agen properti atau ahli hukum untuk memastikan keputusan mereka tepat. Proses ini membantu konsumen memahami berbagai opsi yang tersedia, baik dari sisi harga, lokasi, hingga legalitas transaksi (Akhsya et al., 2021).	Permintaan saran atau arahan dalam pengambilan keputusan pembelian properti, termasuk pertimbangan pemilihan lokasi, perbandingan pilihan unit, rekomendasi agen properti, serta panduan umum sebelum memutuskan untuk membeli.
risiko_dan_disclaimer	Risiko yang terkait dengan pembelian properti dapat mencakup masalah hukum, kualitas bangunan yang buruk, atau kesalahan dalam pengelolaan properti. Oleh karena itu, pengembang properti perlu memberikan disclaimer yang jelas mengenai potensi risiko yang mungkin dihadapi konsumen dalam membeli rumah (Akhsya et al., 2021).	Pertanyaan atau informasi mengenai potensi risiko yang dapat dihadapi konsumen dalam pembelian properti, seperti risiko hukum, ketidaksesuaian kualitas bangunan, keterlambatan serah terima, serta klausul disclaimer yang perlu dipahami sebelum bertransaksi.

Proses pengambilan keputusan dalam pembelian properti melibatkan banyak faktor yang saling mempengaruhi. Konsumen biasanya mempertimbangkan simulasi KPR untuk menilai kemampuan finansial mereka, sementara kualitas

produk dan pelayanan menentukan kepercayaan mereka terhadap pengembang. Proses transaksi dan aspek legalitas juga sangat penting untuk memastikan bahwa pembelian rumah berjalan lancar dan sah. Oleh karena itu, pengembang properti perlu memberikan informasi yang jelas dan transparan tentang setiap aspek yang terkait dengan pembelian properti, mulai dari harga, lokasi, hingga dokumen legal yang diperlukan.

Penelitian oleh Firman (2022) memberikan wawasan lebih lanjut terkait pengaruh bauran pemasaran terhadap keputusan konsumen dalam membeli rumah. Penelitian ini menunjukkan bahwa harga memiliki pengaruh paling dominan terhadap keputusan pembelian rumah. Selain itu, kualitas produk, lokasi, dan promosi juga turut mempengaruhi keputusan konsumen, meskipun dengan tingkat pengaruh yang berbeda. Dalam hal ini, harga yang terjangkau dan lokasi yang strategis menjadi faktor utama yang dipertimbangkan konsumen. Pengembang juga disarankan untuk mempertahankan kualitas produk dan harga yang sesuai dengan kapasitas pasar, serta meningkatkan fasilitas umum di kompleks perumahan untuk mempererat hubungan antar warga (Firman, 2022).

Hal ini bertujuan untuk memberikan wawasan lebih dalam mengenai faktor-faktor yang mempengaruhi keputusan pembelian rumah, serta bagaimana pengembang dapat meningkatkan strategi pemasaran properti agar lebih menarik bagi konsumen.

2.3 Tinjauan Template-Based Data Generation

Metode Template-Based Data Generation menawarkan pendekatan dalam menghasilkan data sintetik untuk pelatihan model, khususnya dalam tugas

klasifikasi intent. Pendekatan ini sangat relevan untuk pengembangan data percakapan chatbot, seperti dalam kasus pembuatan dataset untuk fine-tuning chatbot konsultan properti di Indonesia. Template-Based Data Generation dimulai dengan serangkaian template pertanyaan dasar yang telah ditulis manusia, yang kemudian digunakan untuk menghasilkan berbagai variasi pertanyaan dan jawaban sesuai dengan domain tertentu. Setiap template memiliki placeholder yang dapat diubah untuk menciptakan variasi percakapan yang berbeda-beda.

Dalam konteks pembuatan dataset untuk chatbot, pendekatan ini memanfaatkan kemampuan model bahasa untuk menghasilkan instruksi dan contoh percakapan yang relevan dengan domain tertentu. Misalnya, dalam pembuatan dataset untuk chatbot yang membantu calon pembeli properti, model menghasilkan pertanyaan realistis yang dapat diajukan oleh pengguna seperti "Bagaimana cara menghitung KPR?" atau "Apa saja dokumen yang diperlukan untuk membeli rumah?", lalu memberikan jawaban analitis yang sesuai dengan konteks dan kebutuhan pengguna. Dengan menggunakan Self-Instruct, data dapat diperoleh tanpa memerlukan banyak anotasi manual, yang tentunya lebih hemat waktu dan biaya (Wang et al., 2023).

Salah satu keuntungan utama dari metode ini adalah fleksibilitasnya dalam menghasilkan data yang beragam dan mencakup berbagai jenis pertanyaan dari berbagai profil pengguna, seperti single buyer, pasangan, investor, atau first-time buyer. Hal ini penting untuk menjaga agar model chatbot dapat menangani berbagai jenis percakapan yang mungkin terjadi dalam konteks pembelian properti. Pendekatan ini juga memastikan bahwa data yang dihasilkan lebih beragam dan

tidak terbatas pada pola atau struktur pertanyaan yang serupa, yang sering kali menjadi masalah dalam pengumpulan data manual (Wang et al., 2023; Ding et al., 2024).

Namun, seperti halnya dengan penggunaan data sintetik dalam data augmentation pada umumnya, terdapat tantangan dalam menjaga kualitas dan keakuratan data yang dihasilkan. Penelitian oleh Maheshwari et al. (2023) menunjukkan bahwa data sintetik yang dihasilkan melalui metode template dapat kurang representatif untuk tugas yang lebih kompleks, seperti Named Entity Recognition (NER), yang memerlukan konteks yang lebih mendalam dan pemahaman spesifik terkait entitas dalam percakapan. Oleh karena itu, meskipun Template-Based Data Generation terbukti efektif dalam menghasilkan data untuk tugas-tugas sederhana seperti intent classification, perlu dilakukan penyaringan dan penyesuaian terhadap data yang dihasilkan untuk memastikan relevansi dan kualitasnya sesuai dengan tujuan fine-tuning (Paula et al., 2021).

Selain itu, dengan menggunakan Template-Based Data Generation, kualitas data dapat ditingkatkan secara bertahap dengan memperbaiki dan menyempurnakan template yang ada berdasarkan hasil yang telah dihasilkan sebelumnya. Hal ini memberikan potensi besar bagi pengembangan sistem yang lebih efisien dalam menangani berbagai jenis instruksi tanpa harus bergantung pada manual annotation atau dataset manusia yang mahal dan terbatas (Paula et al., 2021). Dengan memanfaatkan pendekatan ini, proses pembuatan data untuk chatbot konsultan properti dapat diotomatisasi dengan lebih efektif, menghasilkan dataset yang lebih

beragam dan sesuai dengan kebutuhan pengguna dalam berbagai skenario pembelian properti.

2.4 Tinjauan Format Dataset

Format dataset yang digunakan dalam Fine Tuning Large Language Model memiliki pengaruh besar terhadap kemampuan model dalam memahami dan merespons instruksi dengan tepat. Salah satu format yang banyak digunakan dalam fine-tuning LLM adalah format yang melibatkan pasangan intent, instruction, input, dan output. Dalam format ini, intent mengidentifikasi tujuan percakapan, instruction berisi pertanyaan alami yang diajukan oleh pengguna, input memberikan konteks tambahan untuk memperjelas pertanyaan, dan output berupa respons yang terstruktur, analitis, dan edukatif.

Penggunaan format ini terbukti efektif dalam berbagai penelitian terkait fine-tuning LLM. Sebagai contoh, dalam penelitian terkait fine-tuning TinyLLaMA atau dapat disebut sebagai Language Model with Multimodal Applications, format yang serupa digunakan untuk mengembangkan model yang dapat memahami instruksi dan menghasilkan respons yang relevan. Penelitian oleh Wang et al. (2023) menunjukkan bahwa penggunaan format yang jelas dan terstruktur seperti dibawah ini dalam dataset memungkinkan model untuk mempelajari hubungan antara instruksi dan respons dengan lebih baik, meningkatkan kualitas prediksi dan memperluas cakupan tugas yang dapat diselesaikan model.

Instruction: Given an address and city, come up with the zip code.
Input:
Address: 123 Main Street, City: San Francisco

Gambar 2. 1 Contoh Format Dataset

Keberhasilan penggunaan format ini juga didukung oleh bukti dalam berbagai aplikasi praktis. Dalam penelitian oleh Paula et al. (2021), fine-tuning dengan format dataset yang terstruktur memungkinkan chatbot konsultan properti untuk menangani berbagai jenis pertanyaan dari calon pembeli properti dengan lebih efektif. Dengan menggunakan format ini, chatbot dapat diberikan instruksi yang lebih jelas dan spesifik sesuai dengan kebutuhan pengguna, seperti first-time buyers, investor, atau single buyers, yang semuanya membutuhkan jawaban yang terperinci dan relevan.

Selain itu, format ini memungkinkan model untuk belajar lebih cepat karena elemen-elemen dalam dataset yang terstruktur membantu memperjelas konteks dan tujuannya. Misalnya, dalam tugas intent classification, model dapat lebih mudah memahami dan mengklasifikasikan berbagai jenis pertanyaan berdasarkan intent yang sudah ditentukan sebelumnya. Hal ini terbukti efektif dalam meningkatkan akurasi model dalam menangani berbagai pertanyaan pengguna yang memiliki konteks yang berbeda.

Secara keseluruhan, penggunaan format dataset ini dalam fine-tuning TinyLLaMA dan model serupa terbukti tidak hanya meningkatkan kualitas model, tetapi juga meningkatkan efisiensi dalam pelatihan, memungkinkan model untuk

lebih cepat beradaptasi dengan berbagai jenis tugas. Penelitian-penelitian sebelumnya menunjukkan bahwa dataset dengan format yang jelas dan terstruktur sangat membantu dalam pengembangan model yang lebih cerdas dan responsif terhadap instruksi yang diberikan.

2.5 Arsitektur Transformer

Arsitektur Transformer merupakan fondasi utama dari pengembangan berbagai Large Language Models (LLM) modern, termasuk TinyLLaMA, LoRA, dan turunannya. Transformer menggunakan mekanisme self-attention untuk menangkap hubungan antar-token dalam sebuah urutan teks, sehingga lebih efisien dibandingkan arsitektur lama seperti RNN dan LSTM. Inti dari mekanisme ini adalah operasi Scaled Dot-Product Attention, di mana bobot perhatian antara query (Q) dan key (K) digunakan untuk merepresentasikan nilai (V):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

Keterangan:

Q : Matriks query yang digunakan untuk mencari informasi yang relevan

K : Matriks key yang merepresentasikan karakteristik atau penanda dari setiap elemen input.

V : Matriks value berisi informasi aktual yang akan digabungkan berdasarkan bobot perhatian.

T : Transpose Matriks.

d_k : Dimensi vektor *key*, sebagai faktor penskalaan untuk menjaga stabilitas nilai hasil perkalian.

$\text{softmax}()$: Fungsi normalisasi yang mengubah skor kesesuaian menjadi bobot probabilitas.

Dalam konteks fine-tuning, LoRA (Low-Rank Adaptation) diperkenalkan sebagai metode untuk mengadaptasi arsitektur Transformer dengan efisien. LoRA bekerja dengan menambahkan matriks low-rank yang dapat dilatih pada weight matriks Wq dan Wv di lapisan self-attention. Dengan cara ini, sebagian besar bobot

model tetap dibekukan, sementara hanya adaptor kecil yang diperbarui, sehingga menurunkan kebutuhan memori tanpa mengurangi performa (Hu et al., 2021).

Perkembangan LoRA kemudian melahirkan berbagai turunan yang tetap berakar pada arsitektur Transformer. QLoRA menggabungkan LoRA dengan kuantisasi 4-bit sehingga memungkinkan fine-tuning model raksasa seperti LLaMA-65B hanya dengan GPU kelas menengah (Dettmers et al., 2023). Sementara itu, mLoRA berfokus pada efisiensi pelatihan model Transformer skala besar di lingkungan multi-GPU. Pendekatan ini memungkinkan distribusi beban pelatihan secara paralel sehingga model berparameter miliaran dapat diadaptasi tanpa bottleneck signifikan (Ye et al., 2024).

Selain efisiensi pelatihan, beberapa penelitian menyoroti optimisasi struktur Transformer itu sendiri. Sheared LLaMA misalnya, memperkenalkan structured pruning pada model LLaMA untuk memangkas parameter yang kurang signifikan. Meskipun jumlah parameter berkurang, struktur inti Transformer tetap dipertahankan, sehingga model dapat berjalan lebih ringan tanpa kehilangan akurasi secara drastis (Xia et al., 2024).

Konsep Transformer juga diadaptasi dalam pengembangan TinyLLaMA, sebuah model decoder-only yang berjumlah 1,1 miliar parameter. TinyLLaMA mempertahankan prinsip arsitektur Transformer namun dalam ukuran lebih kecil, sehingga dapat dijalankan di perangkat dengan keterbatasan sumber daya tanpa mengorbankan kualitas pemahaman bahasa secara signifikan (Y. Zhang et al., 2024).

Dengan demikian, jelas bahwa arsitektur Transformer menjadi tulang punggung bagi penelitian LLM modern. Inovasi-inovasi seperti LoRA, QLoRA, mLoRA, dan Sheared LLaMA berfokus pada peningkatan efisiensi pelatihan dan inferensi, sementara TinyLLaMA menunjukkan bagaimana varian ringan dari Transformer dapat membuka akses lebih luas bagi penelitian dan implementasi nyata.

2.6 Tinjauan Model TinyLLaMA

Model TinyLLaMA merupakan salah satu pengembangan open-source small language model yang dirancang untuk menjawab kebutuhan akan model bahasa besar (LLM) yang lebih ringan, efisien, dan mudah diakses. Dengan jumlah parameter sekitar 1,1 miliar, TinyLLaMA jauh lebih kecil dibandingkan model populer seperti GPT-3 yang memiliki ratusan miliar parameter (Y. Zhang et al., 2024). Tujuan utama pengembangannya adalah menciptakan model yang tetap kompetitif dalam kualitas, tetapi dapat dijalankan di perangkat dengan keterbatasan sumber daya, seperti laptop dengan GPU menengah atau bahkan server dengan kapasitas terbatas.

Kelebihan TinyLLaMA tidak hanya terletak pada ukurannya yang kecil, tetapi juga pada pendekatan pelatihannya. Meskipun memiliki parameter yang lebih sedikit, TinyLLaMA telah melalui proses pretraining dengan dataset yang cukup besar, sehingga menghasilkan kemampuan pemahaman bahasa alami yang relevan dengan berbagai konteks percakapan. Dalam banyak kasus, TinyLLaMA menunjukkan performa mendekati LLM berukuran lebih besar, terutama pada tugas-tugas percakapan sederhana, klasifikasi teks, dan question answering. Hal ini

membuat TinyLLaMA menjadi alternatif yang praktis untuk penelitian maupun implementasi nyata di sektor bisnis.

Selain TinyLLaMA, sejumlah penelitian lain juga menyoroti keunggulan LLM kecil dalam tugas-tugas spesifik. Studi Tiny Titans: Can Small Language Models Summarize Meetings? menemukan bahwa LLM berparameter kecil mampu menghasilkan ringkasan rapat yang cukup akurat dengan biaya komputasi yang jauh lebih rendah (Fu et al., 2024). Ini menunjukkan bahwa LLM kecil tetap relevan dan dapat digunakan secara efektif dalam aplikasi dengan batasan waktu respons cepat dan keterbatasan energi komputasi. Demikian pula, penelitian GEB: An Efficient Bilingual Language Model with 1.3B Parameters mengembangkan model ringan bilingual dengan dukungan bahasa Inggris dan Mandarin. Model ini menegaskan bahwa LLM berukuran kecil hingga menengah juga dapat diperluas ke ranah multibahasa tanpa mengorbankan efisiensi (Wu et al., 2024).

Tren penelitian yang menekankan pada model ringan seperti TinyLLaMA, Tiny Titans, dan GEB sejalan dengan kebutuhan industri dan akademisi untuk menghadirkan teknologi yang inklusif dan berkelanjutan. Model ini memungkinkan lebih banyak pihak, baik peneliti independen maupun organisasi kecil, untuk mengakses dan mengimplementasikan teknologi LLM tanpa harus bergantung pada infrastruktur komputasi berskala besar. TinyLLaMA dan variannya menunjukkan bahwa inovasi dalam skala kecil tetap dapat memberikan dampak besar, baik untuk aplikasi chatbot, sistem ringkasan otomatis, maupun layanan pelanggan berbasis AI.

2.7 Tinjauan Metode Low-Rank Adaptation

Metode Low-Rank Adaptation (LoRA) diperkenalkan sebagai solusi dalam konteks parameter-efficient fine-tuning (PEFT) yang memungkinkan penyesuaian model bahasa besar (LLM) dengan keterbatasan sumber daya. LoRA bekerja dengan membekukan sebagian besar bobot utama model, kemudian menambahkan matriks low-rank pada lapisan tertentu yang dapat dilatih (Hu et al., 2021). Strategi ini menurunkan jumlah parameter yang dioptimalkan, sehingga kebutuhan memori maupun daya komputasi menjadi jauh lebih rendah. Dengan pendekatan ini, peneliti maupun praktisi dapat memanfaatkan kemampuan LLM besar tanpa harus memiliki perangkat keras berbiaya tinggi seperti GPU server kelas atas.

Secara matematis, misalkan sebuah matriks bobot $W \in \mathbb{R}^{d \times k}$. Pada fine-tuning penuh, semua parameter $d \times k$ akan diperbarui. Namun, pada LoRA, pembaruan bobot dilakukan melalui dekomposisi *low-rank* sebagai berikut:

$$W' = W + \Delta W, \quad \Delta W = A \times B \quad (2.2)$$

Keterangan:

W' : Matriks bobot hasil adaptasi setelah diterapkan metode LoRA

W : Matriks bobot asli dari model pra-latih yang tidak diperbarui secara langsung

ΔW : Matriks perubahan bobot yang dihasilkan oleh mekanisme LoRA

$A \in \mathbb{R}^{(d \times r)}$: Matriks *low-rank* yang memproyeksikan fitur ke ruang berdimensi lebih kecil

$B \in \mathbb{R}^{(r \times k)}$: Matriks adaptasi yang memproyeksikan kembali hasil ke dimensi semula

r : Nilai *rank* rendah pada LoRA yang mengontrol jumlah parameter tambahan

Jumlah parameter yang perlu dilatih pada LoRA menjadi:

$$\text{Parameter}_{\text{LoRA}} = r \times (d + k) \quad (2.3)$$

Keterangan:

$\text{Parameter}_{\text{LoRA}}$: Jumlah parameter tambahan yang dihasilkan oleh mekanisme LoRA

r : Nilai *rank* rendah (*low-rank*) yang digunakan pada matriks adaptasi LoRA

d : Dimensi vektor masukan atau keluaran dari layer yang diadaptasi

k : Dimensi vektor keluaran atau jumlah fitur pada layer terkait

Sedangkan jumlah parameter pada fine-tuning penuh adalah:

$$\text{Parameter Full} = d \times k \quad (2.4)$$

Keterangan:

Parameter Full : Jumlah parameter tambahan yang dihasilkan oleh mekanisme LoRA

d : Dimensi vektor masukan atau keluaran dari layer yang diadaptasi

k : Dimensi vektor keluaran atau jumlah fitur pada layer terkait

LoRA kemudian berkembang menjadi fondasi bagi berbagai variasi teknik adaptasi. Salah satunya adalah RoRA (Rank-regularized Low-rank Adaptation) yang memperkenalkan regularisasi pada rank untuk mengatasi ketidakstabilan yang sering muncul dalam LoRA standar. Dengan adanya regularisasi ini, RoRA dapat meningkatkan stabilitas pelatihan sekaligus menjaga kualitas hasil keluaran model (Liu et al., 2025). Di sisi lain, mLoRA (multi-GPU LoRA) dirancang untuk mendukung pelatihan skala besar dengan memanfaatkan kemampuan multi-GPU. Dengan demikian, model berukuran besar seperti LLaMA-13B atau bahkan lebih besar dapat dilatih dengan efisien tanpa harus melakukan kompromi pada performa (Ye et al., 2024).

Selain itu, inovasi penting lainnya adalah QLoRA (Quantized LoRA) yang menggabungkan prinsip LoRA dengan teknik kuantisasi 4-bit. Dengan kuantisasi ini, memori yang dibutuhkan untuk menjalankan model menjadi jauh lebih kecil, sehingga memungkinkan fine-tuning pada model hingga 65 miliar parameter hanya dengan menggunakan satu GPU kelas menengah (Dettmers et al., 2023). QLoRA menjadi terobosan signifikan karena membuka akses bagi lebih banyak pihak untuk melatih model ultra-besar tanpa keterbatasan perangkat keras.

Meskipun bukan bagian langsung dari keluarga LoRA, penelitian seperti Sheared LLaMA juga relevan karena fokus pada peningkatan efisiensi model

melalui structured pruning. Teknik ini memangkas parameter model secara sistematis, sehingga ukuran model dapat diperkecil tanpa mengorbankan akurasi secara signifikan (Xia et al., 2024). Pendekatan ini memperlihatkan bahwa arah pengembangan LLM modern tidak hanya berfokus pada peningkatan kemampuan, tetapi juga pada pengurangan beban komputasi dan pemanfaatan sumber daya secara lebih hemat.

Secara keseluruhan, LoRA dan variasinya telah menjadi elemen kunci dalam perkembangan penelitian LLM. Penerapan metode ini memungkinkan berbagai eksperimen dilakukan dengan biaya yang lebih rendah, serta mendorong integrasi LLM ke berbagai aplikasi nyata seperti chatbot, sistem rekomendasi, layanan pelanggan, dan pendidikan berbasis AI. Inovasi-inovasi tersebut mempertegas bahwa masa depan pengembangan model bahasa tidak hanya ditentukan oleh ukuran parameter, tetapi juga oleh efisiensi dan keterjangkauan dalam proses adaptasi.

2.8 Metode Evaluasi Model dengan BLEU

Dalam pengembangan chatbot berbasis NLP, aspek evaluasi menjadi krusial karena menentukan apakah keluaran model benar-benar sesuai dengan ekspektasi pengguna maupun standar akademik. Ada sebuah metode evaluasi yang banyak digunakan dalam penelitian NLP modern adalah BLEU (Bilingual Evaluation Understudy).

BLEU merupakan metrik berbasis precision yang menilai kesesuaian keluaran model dengan teks referensi melalui tingkat kemiripan n-gram. Secara umum, semakin banyak n-gram (kata atau frasa pendek) dari keluaran model yang

cocok dengan referensi, semakin tinggi skor BLEU yang diperoleh. BLEU juga memperhitungkan panjang keluaran agar tidak terjadi bias terhadap jawaban pendek melalui faktor penalti yang disebut brevity penalty (BP). Rumus BLEU dituliskan sebagai berikut:

$$\text{BLEU} = \text{BP} \times \left(\sum_{n=1}^N \omega_n \log p_n \right) \quad (2.5)$$

Keterangan:

BLEU : Nilai evaluasi yang mengukur tingkat kemiripan antar teks

BP : Brevity Penalty yang digunakan untuk mengurangi nilai BLEU

N : Orde maksimum *n-gram* untuk menentukan tingkat detail kecocokan urutan kata

ω_n : Bobot untuk *n-gram* ke-*n* yang berfungsi mengatur kontribusi setiap orde *n-gram*

p_n : Modified *n-gram* precision yang menunjukkan proporsi *n-gram* keluaran model

Sejumlah penelitian yang dijadikan rujukan juga menempatkan BLEU sebagai metrik utama. Misalnya, pada penelitian Fu et al mengevaluasi efektivitas LLM kecil untuk ringkasan rapat dengan menggunakan BLEU sebagai tolok ukur kuantitatif (Fu et al., 2024). Hasil penelitian tersebut menunjukkan bahwa meskipun model kecil seperti TinyLLaMA tidak dapat sepenuhnya menyamai performa LLM besar, mereka tetap mampu menghasilkan ringkasan dengan skor BLEU yang cukup kompetitif, sehingga lebih hemat biaya komputasi tanpa kehilangan relevansi.

Penelitian A Comparative Study of Fine-Tuning Strategies for Large Language Models juga memanfaatkan BLEU untuk membandingkan strategi fine-tuning seperti QLoRA, RAFT, dan RLHF (Meyer et al., 2024). Dalam penelitian ini, BLEU digunakan untuk mengukur kelengkapan informasi dalam ringkasan atau jawaban yang dihasilkan. Hasilnya memperlihatkan bahwa QLoRA relatif lebih

efisien, namun RAFT dan RLHF cenderung memberikan skor yang lebih baik dalam konteks semantik dan kualitas jawaban berbasis preferensi manusia.

Lebih jauh lagi, penelitian terkait LoRA dan variasinya, seperti RoRA (Liu et al., 2025) dan mLoRA (Ye et al., 2024), meskipun lebih berfokus pada efisiensi pelatihan, tetap menggunakan BLEU sebagai alat validasi standar. Hal ini karena metrik tersebut telah diakui secara luas dalam bidang NLP sebagai baseline evaluation metrics yang dapat dibandingkan lintas penelitian.

Meskipun BLEU memiliki peran penting, namun tetap memiliki keterbatasan. BLEU cenderung menilai overlap leksikal tanpa memperhitungkan sinonim atau parafrasa, sehingga dapat meremehkan keluaran yang semantik-nya benar namun berbeda kata.

Untuk mengatasi hal tersebut, beberapa penelitian lanjutan menambahkan evaluasi berbasis manusia atau menggunakan metrik semantik modern seperti BERTScore atau evaluasi berbasis GPT-4. Sebagai contoh, (Gnewuch et al., 2024) dalam studi hybrid chatbot menekankan pentingnya evaluasi manusia untuk menangkap aspek user satisfaction, sedangkan Pandya & Holia (2023) dalam penelitian berbasis LangChain menunjukkan bahwa metrik otomatis perlu dilengkapi dengan pengukuran berbasis kualitas interaksi nyata.

Dengan demikian, BLEU tetap relevan sebagai fondasi evaluasi chatbot berbasis LLM. Meskipun terbatas metrik ini mampu memberikan tolok ukur kuantitatif yang objektif. Penggunaannya dalam berbagai penelitian TinyLLaMA, LoRA, QLoRA, dan Tiny Titans memperlihatkan bahwa BLEU bukan hanya standar historis, tetapi juga alat yang terus dipakai dalam menilai kualitas keluaran

model bahasa modern. Kombinasi dengan evaluasi semantik dan human evaluation menjadi strategi ideal agar hasil penelitian lebih komprehensif dan mendalam.

2.9 Metode Uji Implementasi Model

Uji implementasi model pada sistem chatbot umumnya dilakukan melalui instrumen kuesioner yang mengukur beberapa dimensi utama, yaitu kualitas jawaban dan kepuasan pengguna. Titoni et al. (2024) membuktikan bahwa responsivitas chatbot berpengaruh signifikan terhadap pengalaman pengguna, dengan hasil uji hipotesis menunjukkan pengaruh positif responsivitas terhadap extrinsic value, intrinsic value, dan online convenience. Senada dengan itu, Al Jupri & Fasa (2025) menemukan bahwa kesesuaian dan akurasi jawaban chatbot terhadap pertanyaan pengguna merupakan faktor penting dalam meningkatkan kepuasan pelanggan, sehingga pengukuran akurasi dan kelengkapan jawaban menjadi indikator yang relevan dalam instrumen evaluasi.

Dari sisi kualitas jawaban, Soetiyono et al. (2024) mengoperasionalkan variabel chatbot melalui indikator kejelasan, ketepatan waktu atau responsiveness, dan kemampuan memahami bahasa alami pengguna, dengan hasil penelitian membuktikan pengaruh signifikan chatbot terhadap kepuasan pelanggan. Manik et al. (2026) memperkuat hal ini dengan menemukan bahwa penilaian positif pengguna terhadap chatbot didominasi oleh aspek kemudahan penggunaan, dan akurasi jawaban, sementara keluhan utama berkaitan dengan ketidakakuratan jawaban dan ketidakmampuan chatbot memahami konteks yang kompleks. Kedua temuan ini menegaskan bahwa kelengkapan, akurasi, dan kemudahan pemahaman

jawaban merupakan dimensi yang perlu diukur secara eksplisit dalam instrumen evaluasi.

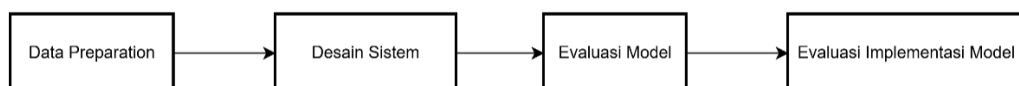
Terkait kepuasan pengguna dan perbandingan antara chatbot dengan agen manusia, Al Jupri & Fasa (2025) menyatakan bahwa meskipun chatbot terbukti meningkatkan efisiensi layanan, chatbot masih memiliki keterbatasan dalam menangani pertanyaan kompleks sehingga integrasi dengan agen manusia tetap diperlukan. Perbandingan langsung antara kedua sistem dari perspektif pengguna menjadi relevan untuk mengidentifikasi keunggulan dan kelemahan masing-masing. Dalam hal instrumen pengukuran, Soetiyono et al. (2024) menggunakan skala Likert 1–5 untuk mengkuantifikasi persepsi pengguna, sementara Manik et al. (2026) menggabungkan pertanyaan terstruktur berskala Likert dengan umpan balik kualitatif terbuka, pendekatan yang memungkinkan evaluasi yang lebih komprehensif terhadap pengalaman pengguna secara keseluruhan.

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Penelitian

Penelitian ini dimulai pada tahap Data Preparation, yang melibatkan beberapa langkah terstruktur. Proses dimulai dengan pemilihan topik intent yang relevan untuk penelitian ini. Setelah itu, format dataset ditentukan agar proses pelatihan bisa berjalan lebih efektif. Selanjutnya, dibuat prompt yang akan digunakan untuk menghasilkan data percakapan. Data kemudian dihasilkan menggunakan model GPT berdasarkan prompt yang telah dibuat. Selanjutnya kualitas dan relevansi dataset akan dinilai oleh agen properti melalui kuesioner. Dataset yang telah lolos validasi kemudian dipersiapkan untuk analisis lebih lanjut.



Gambar 3. 1 Desain Penelitian

Tahap berikutnya adalah Desain Sistem, di mana arsitektur chatbot ditetapkan dengan memilih model TinyLLaMA, menetapkan pendekatan fine-tuning LoRA, serta menentukan konfigurasi pelatihan yang meliputi jumlah rank, lora alpha, learning rate, dan jumlah epoch agar selaras dengan tujuan eksperimen. Setelah rancangan siap, penelitian masuk ke Evaluasi Model. Model yang sudah terlatih diuji menggunakan metrik BLEU-4 untuk melihat sejauh mana model mampu menghasilkan jawaban yang relevan, akurat, dan konsisten dengan konteks.

Hasil evaluasi ini menjadi dasar penyusunan kesimpulan serta rekomendasi pengembangan penelitian selanjutnya.

Tahap terakhir pada penelitian ini menilai dampak penerapan chatbot terhadap alur layanan agensi. Pengukuran dilakukan melalui uji coba dengan membandingkan penggunaan chatbot dan interaksi manual melalui WhatsApp untuk melihat peningkatan kinerja. Pengukuran ini mencakup kualitas respons dengan menggunakan umpan balik dari pengguna terhadap kualitas layanan. Hasilnya dibandingkan dengan kondisi dasar untuk menilai sejauh mana sistem baru berkontribusi pada optimalisasi layanan dan menjadi landasan untuk rekomendasi pengembangan berikutnya. Hasilnya dibandingkan terhadap kondisi dasar untuk menilai sejauh mana sistem berkontribusi pada optimalisasi layanan dan menjadi landasan rekomendasi pengembangan berikutnya.

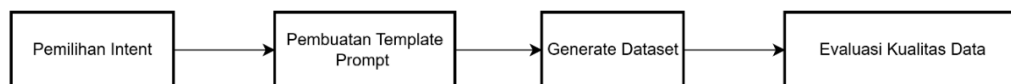
3.2 Data Preparation

Bagian ini menjelaskan proses pembuatan data sintetis yang menjadi dasar pelatihan model dalam penelitian ini. Proses dimulai dengan pemilihan topik intent yang relevan dengan keputusan pembelian properti. Setelah intent dipilih, langkah selanjutnya adalah menentukan format data pelatihan dan membuat template prompt untuk generate data yang sesuai dengan kebutuhan fine-tuning. Data percakapan kemudian dihasilkan menggunakan metode Template-Based data Generation, di mana nantinya model LLM akan digunakan untuk menghasilkan pasangan instruksi dan respons dalam format JSON berdasarkan prompt yang telah dibuat. Proses ini memastikan bahwa dataset yang dihasilkan memiliki kualitas dan relevansi yang memadai untuk mendukung pembelajaran model TinyLLaMA

dengan metode LoRA. Dataset yang dihasilkan selanjutnya akan dievaluasi secara kualitatif untuk memastikan kesesuaian dan keberagaman data yang digunakan dalam pelatihan.

3.2.1 *Template-Based Data Generation*

Metode Template-Based Data Generation merupakan pendekatan yang digunakan untuk menghasilkan dataset sintetik yang digunakan dalam fine-tuning chatbot. Proses ini dimulai dengan pemilihan intent atau topik dengan domain yang ingin dilayani oleh chatbot. Intent ini berfungsi sebagai dasar untuk membuat berbagai skenario percakapan yang dapat dijawab oleh chatbot. Pemilihan intent sangat penting karena menentukan jenis-jenis pertanyaan yang akan dimiliki chatbot dan memastikan relevansi dengan konteks yang ada.



Gambar 3. 2 Alur Template-Based Data Generation

Setelah pemilihan intent, langkah selanjutnya adalah pembuatan template prompt untuk mengatur proses generasi data dan menentukan format data sehingga dapat menghasilkan pertanyaan relevan. Nantinya template ini digunakan untuk menghasilkan dataset sintetik yang terstruktur dan sistematis, yang selanjutnya digunakan dalam pelatihan model chatbot. Terakhir, kualitas dataset dievaluasi melalui questioner untuk memastikan kecocokan data dengan skenario nyata.

A. Pemilihan Intent

Pemilihan intent adalah langkah pertama dalam pembuatan dataset menggunakan pendekatan template. Intent ini menggambarkan tujuan dari percakapan yang akan dijalankan oleh chatbot. Dalam konteks chatbot konsultan properti, intent yang dipilih harus sesuai dengan kebutuhan calon pembeli, seperti informasi mengenai KPR, legalitas dokumen, spesifikasi properti, dan lainnya. Dengan menentukan intent yang jelas, proses pembuatan template akan lebih terarah dan sesuai dengan kebutuhan pengguna.

Dataset percakapan yang digunakan mencakup tujuh topik atau intent yang merepresentasikan berbagai konteks interaksi antara agen dan pelanggan di sektor properti. Setiap topik menggambarkan jenis informasi yang paling sering muncul dalam layanan pelanggan dan membantu model memahami maksud pengguna secara lebih spesifik.

Tabel 3. 1 List Intent Dataset

Intent	Contoh Pertanyaan
kpr_simulasi	“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 saya bisa cicil atau tidak ya?”
legal_dokumen	“Saya tertarik beli rumah second, tapi masih bingung dokumen apa saja yang wajib saya cek supaya aman secara hukum.”
biaya_dan_pajak	“Kalau saya beli rumah Rp410 juta lewat KPR, biaya tambahan di luar DP biasanya apa saja dan kisarannya berapa?”
spesifikasi_properti	“Saya lihat rumah tipe 45/90, tapi kurang paham arti angka itu dan apa implikasinya buat kenyamanan tinggal.”
proses_transaksi	“Sebagai pembeli pertama kali, saya ingin tahu urutan proses beli rumah lewat KPR dari awal sampai akad.”
konsultasi_keputusan	“Saya dan istri ragu, lebih baik beli rumah sekarang atau menunggu tabungan bertambah supaya DP lebih besar?”
risiko_dan_disclaimer	“Apa saja risiko terbesar kalau saya ambil KPR jangka panjang sampai 25 tahun?”

Ketujuh topik tersebut memastikan variasi konteks yang memadai untuk melatih model agar mampu memberikan respons yang relevan terhadap berbagai situasi dalam percakapan layanan properti.

B. Penentuan Format Data

Penentuan format data sintetis adalah langkah penting dalam pembuatan dataset untuk fine-tuning model chatbot, yang bertujuan untuk memastikan data yang dihasilkan sesuai dengan tujuan pelatihan. Dalam proses ini, setiap entri data dibangun dengan atribut yang jelas dan terstruktur untuk memastikan model dapat memproses dan memahami setiap komponen informasi dengan tepat. Format data yang digunakan untuk chatbot konsultan properti terdiri dari beberapa atribut utama, yaitu intent, instruction, input, dan output. Setiap atribut memiliki peran spesifik yang mendukung tujuan chatbot untuk memberikan respons yang relevan dan akurat sesuai dengan konteks pertanyaan yang diajukan oleh pengguna.

Tabel 3. 2 Format Dataset

Atribut	Tipe Data	Deskripsi	Contoh
intent	String	Intent tiap entri data	"biaya_dan_pajak"
instruction	String	Pertanyaan dari customer (hasil kurasi/augmentasi)	"Kalau saya beli rumah Rp410 juta lewat KPR, biaya tambahan di luar DP biasanya apa saja dan kisarannya berapa?"
input	String	(Opsional) konteks tambahan, mis. ringkas spesifikasi/unit	""
output	String	Penjelasan chatbot yang diharapkan	"Selain DP, ada sejumlah biaya lain yang sering membuat pembeli kaget jika tidak dipersiapkan. Saya jelaskan secara bertahap dengan asumsi harga rumah Rp410 juta.\n\n1) **Biaya notaris & akta**: Untuk

			AJB, APHT, dan pengikatan kredit, biasanya total bisa 1%–2% dari nilai transaksi.\n2) **BPHTB**: Pajak pembeli sebesar 5% dari nilai kena pajak setelah dikurangi NJOPTKP daerah.\n3) **Biaya provisi bank**: Umumnya sekitar 1% dari plafon pinjaman.\n4) **Asuransi jiwa & kebakaran**: Besarnya tergantung usia dan nilai kredit, dibayar di awal.\n5) **Biaya administrasi & appraisal**: Nominal tetap atau persentase kecil dari kredit.\n\nTotal tambahan sering berada di kisaran 6%–10% dari harga rumah. Angka ini bisa berbeda tiap daerah dan bank, jadi tetap perlu konfirmasi detail sebelum transaksi."
--	--	--	---

Tabel 3.2 menunjukkan format dataset yang digunakan dalam proses pembuatan data untuk fine-tuning chatbot. Atribut pertama adalah intent, yang menunjukkan kategori atau tujuan dari setiap percakapan, misalnya "biaya_dan_pajak", yang mengindikasikan bahwa percakapan tersebut berfokus pada biaya terkait pembelian properti. Selanjutnya, atribut instruction berisi pertanyaan atau instruksi yang diberikan oleh customer, yang mencerminkan pertanyaan nyata yang bisa diajukan, seperti "Kalau saya beli rumah Rp410 juta lewat KPR, biaya tambahan di luar DP biasanya apa saja dan kisarannya berapa?". Atribut input adalah konteks tambahan yang bersifat opsional, dan memberikan informasi lebih lanjut yang relevan untuk menjawab pertanyaan, meskipun tidak selalu diperlukan. Terakhir, atribut output berisi jawaban dari chatbot yang diharapkan, yang memberikan penjelasan rinci dan terstruktur kepada pengguna

sesuai dengan pertanyaan yang diajukan, seperti penjelasan mengenai biaya-biaya tambahan dalam pembelian rumah. Struktur ini memastikan data yang dihasilkan memiliki informasi yang lengkap dan relevan, yang akan digunakan untuk melatih chatbot dalam merespons berbagai jenis pertanyaan dengan cara yang edukatif dan informatif.

C. Penentuan Template Prompt

Penentuan template prompt merupakan langkah penting dalam proses pembuatan dataset untuk fine-tuning chatbot, karena template ini menjadi dasar dalam menghasilkan variasi percakapan yang realistis dan relevan dengan konteks pengguna. Template prompt yang digunakan bertujuan untuk memastikan bahwa data yang dihasilkan sesuai dengan kebutuhan spesifik chatbot konsultan properti, yang dapat menangani beragam pertanyaan dari calon pembeli properti dengan berbagai profil, seperti first-time buyer, pasangan, investor, atau single buyer.

Tabel 3. 3 Prompt Generator Dataset

Bagian	Prompt
Peran	Peran Anda adalah generator dataset untuk fine-tuning chatbot konsultan properti di Indonesia.
Tugas Utama	Tugas Anda adalah menghasilkan data percakapan unik (tidak boleh mengulang kalimat, struktur, atau contoh sebelumnya) dalam format JSONL untuk keperluan supervised fine-tuning (SFT), dengan ketentuan berikut:
Format Wajib	FORMAT WAJIB: Setiap data HARUS berbentuk: <pre>{ "intent": "<salah satu intent yang diizinkan>", "instruction": "<pertanyaan alami dan realistis dari calon pembeli properti>", "input": ""</pre>

	"output": "<jawaban panjang, analitis, terstruktur, dan edukatif dari konsultan properti>" }
Intent yang Diizinkan	#INTENT YANG BOLEH DIGUNAKAN: kpr_simulasi legal_dokumen biaya_dan_pajak spesifikasi_properti proses_transaksi konsultasi_keputusan risiko_dan_disclaimer
Aturan Keunikan	#ATURAN PENTING: 1. Setiap data HARUS UNIK, tidak boleh: - Mengulang contoh angka yang sama - Mengulang struktur kalimat - Mengulang skenario kasus
Variasi User	2. Pertanyaan user harus realistis & bervariasi: - Single buyer - Pasangan - Investor - First-time buyer
Ketentuan Jawaban (Output)	3. Jawaban assistant HARUS: - Minimal 120 kata - Mengandung penjelasan bertahap atau poin bernomor - Menyebut asumsi (jika relevan) - Menyertakan catatan risiko atau disclaimer ringan - DILARANG hard selling, CTA marketing, atau menyebut nama developer tertentu Gunakan bahasa profesional tapi mudah dipahami
Output Akhir	OUTPUT: > Hanya tampilkan JSON (tanpa penjelasan tambahan) > 1 baris = 1 JSON object (JSONL) > Buat sebanyak N data sesuai permintaan

Template prompt yang digunakan dalam pengembangan dataset ini mengarah pada pembuatan data percakapan yang unik dan tidak mengulang kalimat atau struktur yang sama, untuk menjaga keberagaman dalam dataset. Dalam template prompt, setiap data percakapan harus memuat atribut-atribut tertentu, seperti intent (tujuan percakapan), instruction (pertanyaan dari calon pembeli),

input (konteks tambahan jika diperlukan), dan output (jawaban chatbot yang edukatif dan analitis). Misalnya, dalam template prompt yang digunakan untuk pembuatan dataset chatbot konsultan properti, pertanyaan yang muncul bisa beragam, seperti pertanyaan mengenai KPR, biaya tambahan, spesifikasi properti, atau legalitas dokumen.

Selain itu, template prompt ini memiliki aturan ketat terkait variasi data. Setiap data yang dihasilkan harus unik, baik dalam struktur kalimat maupun dalam skenario kasus. Hal ini dilakukan untuk memastikan bahwa model chatbot dapat belajar menangani berbagai skenario percakapan yang mungkin terjadi dalam konteks pembelian properti. Dengan aturan tersebut, template prompt membantu memastikan bahwa model dapat memberikan jawaban yang bervariasi dan relevan sesuai dengan profil dan kebutuhan pengguna yang berbeda-beda. Sebagai contoh, chatbot akan memberikan penjelasan tentang biaya tambahan dalam pembelian rumah untuk seorang first-time buyer dibandingkan dengan investor yang sudah lebih berpengalaman. Secara keseluruhan, template prompt ini berfungsi sebagai panduan dalam menciptakan data percakapan yang beragam dan sesuai dengan berbagai konteks.

D. Generate Dataset

Setelah template prompt disiapkan, langkah berikutnya adalah menghasilkan dataset sintetik dengan target sebanyak 3024 data. Proses ini menggunakan model GPT untuk menghasilkan beragam variasi pertanyaan dan jawaban berdasarkan template yang telah disusun. Dengan memanfaatkan kemampuan model untuk memahami konteks dan menghasilkan teks yang relevan,

dataset ini berfokus pada skenario percakapan yang realistis dalam konteks pembelian properti. Setiap data percakapan yang dihasilkan akan meliputi intent yang telah ditentukan sebelumnya, diikuti dengan instruksi yang mencerminkan pertanyaan alami dari calon pembeli properti, serta jawaban yang analitis dan terstruktur dengan jelas. Berikut adalah contoh gambaran dari generate dataset:

Tabel 3. 4 Hasil Generate 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."

Dengan menggunakan GPT, proses pembuatan dataset sintetik menjadi lebih efisien, karena model ini dapat menghasilkan data dalam jumlah besar, tanpa memerlukan banyak anotasi manual.

3.2.2 Evaluasi Kualitas Dataset dengan Quesioner

Untuk memastikan kualitas dataset yang dihasilkan oleh model chatbot properti, dilakukan evaluasi menggunakan quesioner yang dirancang untuk mengukur berbagai aspek penting dari dataset sintetik. Evaluasi ini melibatkan evaluasi manusia sebagai langkah kunci dalam menilai relevansi, keragaman, keakuratan, dan kemudahan pemahaman data yang dihasilkan.

Evaluasi kualitas dataset akan dibagi menjadi beberapa tahap yang meliputi pembagian dataset berdasarkan kategori intent, serta penilaian konteks dan bahasa jawaban chatbot. Evaluasi ini akan dilakukan oleh beberapa evaluator dengan latar belakang dalam bidang properti, sehingga mereka dapat memberikan penilaian yang lebih mendalam terkait kualitas dataset. Berikut dibawah ini merupakan alur proses evaluasinya:

A. Pembagian Subset Dataset

Dataset akan dibagi berdasarkan kategori intent yang relevan dengan konsultasi properti. Setiap kategori intent akan berisi 2-3 data representatif untuk evaluasi. Beberapa kategori yang akan dievaluasi antara lain KPR Simulasi, Legal Dokumen, Biaya dan Pajak, Proses Transaksi, Konsultasi Keputusan, Risiko dan Disclaimer, dan Spesifikasi Property.

B. Kuesioner Evaluasi

Evaluator akan diminta untuk menjawab pertanyaan yang dibagi menjadi dua bagian, yaitu identifikasi intent dan variasi jawaban serta penilaian realitas dan kelayakan jawaban. Identifikasi Intent dan Variasi Jawaban Evaluator akan menilai

seberapa beragam pertanyaan yang diajukan dan apakah mereka bervariasi dalam konteks yang sesuai untuk masing-masing profil pengguna. Penilaian Realitas dan Kelayakan Jawaban Evaluator akan diberikan satu pertanyaan dari dataset dan diminta untuk menilai apakah jawaban chatbot terasa alami, informasi yang relevan, dan mudah dipahami oleh calon pembeli properti.

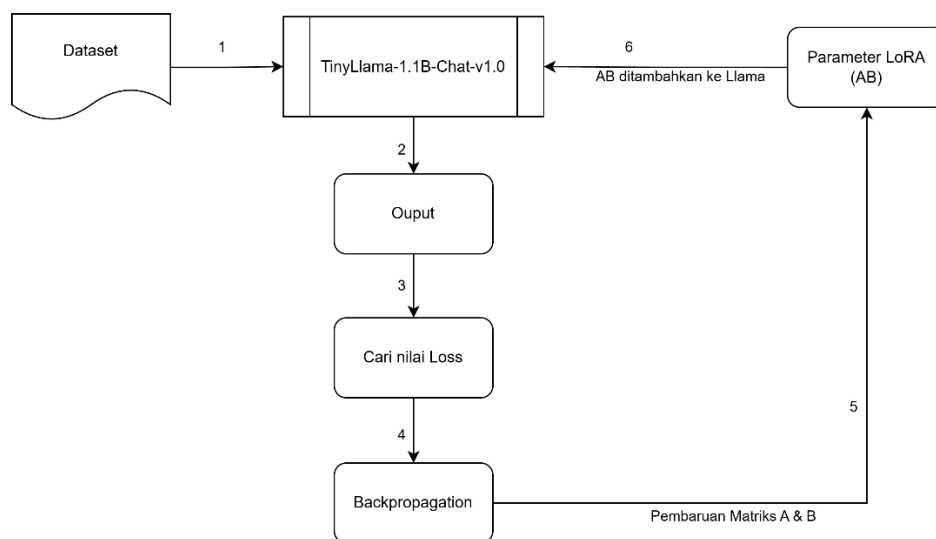
C. Simpulan dan Perbaikan Dataset

Berdasarkan hasil kuesioner data yang dianggap kurang relevan atau tidak akurat akan diperbaiki. Evaluasi ini akan membantu dalam meningkatkan kualitas dataset yang dihasilkan dan menghasilkan data yang lebih alami, relevan, dan mudah dipahami.

3.3 Desain Sistem

Desain sistem pada Gambar 3.3 berfungsi untuk memetakan alur kerja dari dataset hingga menghasilkan output berupa LoRA parameter yaitu matriks A dan B, sehingga setiap tahap terlihat jelas. Alur sistem dimulai dari dataset yang terlebih dahulu dibagi menjadi dua bagian, yaitu training data dan evaluation data. Data yang telah dipisahkan kemudian digunakan dalam proses pelatihan model. Selanjutnya, dilakukan penentuan parameter LoRA berupa dua matriks, yaitu matriks A dan B. Dimensi kedua matriks ini ditentukan berdasarkan nilai rank yang dipilih, di mana matriks A berukuran $r \times d_{\text{model}}$ dan matriks B berukuran $d_{\text{model}} \times r$, dengan r menyatakan rank dan d_{model} merupakan dimensi dari model. Selanjutnya, kedua matriks tersebut diinisialisasi dengan nilai awal secara acak. Parameter LoRA ini kemudian dijumlahkan ke dalam bobot model TinyLLaMA-

1.1B-Chat-v1.0, sehingga model dapat mempelajari penyesuaian bobot tanpa harus memperbarui seluruh parameter utama. Pada proses pelatihan, data dari dataset dimasukkan ke dalam model untuk menghasilkan output. Output tersebut digunakan untuk menghitung nilai loss, yang merepresentasikan tingkat kesalahan prediksi model. Melalui proses backpropagation, gradien dihitung dan digunakan untuk memperbarui parameter. Dalam pendekatan LoRA matriks A dan B diperbarui selama training, sementara parameter asli model tetap dibekukan atau dibiarkan.



Gambar 3. 3 Desain Sistem

3.3.1 Dataset

Dataset sintetis yang digunakan dalam penelitian ini dibuat menggunakan metode Template-Based Data Generation, yang menghasilkan data percakapan terkait pembelian properti, simulasi KPR, legalitas dokumen, biaya, dan pajak. Setiap kategori intent, seperti "simulasi KPR", "legal dokumen", dan "biaya dan pajak", dipilih untuk mencakup variasi pertanyaan yang relevan dengan konteks

pembelian properti. Dengan menggunakan template, percakapan yang dihasilkan terdiri dari pasangan instruksi dan respons yang menggambarkan interaksi yang mungkin terjadi antara pengguna dan chatbot. Proses ini memastikan bahwa dataset yang dihasilkan memenuhi kebutuhan pelatihan model dengan variasi yang cukup, serta kualitas dan relevansi yang sesuai dengan konteks yang diinginkan. Serta selanjutnya dilakukan proses evaluasi dari ahli di bidang agensi property.

Pada tahap selanjutnya adalah membagi dataset menjadi dua bagian, yakni 90% untuk training dan 10% untuk evaluasi. Dalam pembagian ini dataset training diacak dan diambil secara acak, sedangkan dataset evaluasi dibatasi pada sekitar 300 data untuk memastikan evaluasi yang representatif. Pemisahan dilakukan per percakapan untuk menghindari overlap, memastikan bahwa satu percakapan hanya digunakan dalam satu set training atau evaluasi, sehingga hasil evaluasi tetap adil dan tidak terjadi kebocoran informasi antar kedua set.

3.3.2 Low Rank Adaptation

Low-Rank Adaptation merupakan teknik fine-tuning yang hanya menambahkan sejumlah kecil parameter tambahan ke dalam LLM yang telah dilatih sebelumnya. Pada pendekatan ini, bobot asli model $W \in R^{d \times k}$ tetap beku, sementara dua matriks tambahan $A \in R^{r \times d_{\text{model}}}$ dan $B \in R^{d_{\text{model}} \times r}$ dilatih untuk membentuk representasi rank-rendah dari update parameter. Dengan demikian, output akhir dihitung sebagai:

$$W' = W + (A \cdot B) \quad (3.1)$$

Keterangan :

$x \in R^k$: vektor input
 $W \in R^{d \times k}$: matriks bobot model asli
 $A \in R^{r \times d}$: matriks down-rank projection LoRA
 $B \in R^{d \times r}$: matriks up-rank projection LoRA

$r \ll d, k$: low rank LoRA
 $AB \in R^{d \times k}$: matriks update low rank yang training selama fine-tuning
 W' : output akhir layer setelah penyesuaian LoRA

Misalnya jika kita asumsikan dimensi $d = 4$, $k = 3$, dan rank $r = 2$, maka update low rank akan seperti $AB \in R^{4 \times 3}$. Perhitungan manual sederhana: jika A dan B masing-masing memiliki 8 dan 6 parameter, maka total parameter tambahan hanya 14, jauh lebih kecil dibandingkan dengan full fine-tuning pada W yang memiliki 12 parameter. Skala efisiensi ini diperkuat dengan konfigurasi LoRA hyperparameters, yaitu:

1. r (*rank*) : menentukan ukuran dekomposisi rank-rendah.
2. $Lora\ alpha$: faktor skala untuk menyeimbangkan kontribusi update LoRA terhadap bobot asli.
3. $Lora\ dropout$: regularisasi untuk mencegah overfitting pada parameter tambahan.

Pada implementasi penelitian ini, adaptor LoRA disisipkan pada lapisan attention model TinyLLaMA, khususnya pada proyeksi query dan value. pemilihan lapisan ini mengacu pada temuan Hu et al. (2021) bahwa kapasitas generalisasi model sebagian besar berada pada blok perhatian, sehingga menyesuaikan bagian ini sudah cukup untuk mengubah gaya respons tanpa harus memperbarui seluruh bobot model.

Secara konkret, penelitian ini menempatkan adaptor LoRA pada proyeksi W_Q dan W_V . Alih-alih memperbarui seluruh matriks W_Q dan W_V secara langsung, LoRA memperkenalkan dua pasang matriks berukuran kecil, yaitu matriks A dan matriks B , yang hasil perkaliannya dijumlahkan ke bobot pretrained asli. Untuk

proyeksi query, pembaruan bobot diformulasikan sebagaimana ditunjukkan pada

Persamaan 3.2:

$$W'_Q = W_Q + B_Q A_Q \quad (3.2)$$

Keterangan :

$W_Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$: matriks bobot pretrained untuk proyeksi query pada mekanisme self-attention

$W'_Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$: matriks bobot query setelah dilakukan fine-tuning

$A_Q \in \mathbb{R}^{r \times d_k}$: matriks down projection pada LoRA

$B_Q \in \mathbb{R}^{d_{\text{model}} \times r}$: matriks upper projection pada LoRA

$r \ll d_{\text{model}}, d_k$: rank LoRA yang membuat fine-tuning jadi ringan

$B_Q A_Q$: matriks low-rank update yang dipelajari saat fine-tuning

Di mana W_Q adalah matriks bobot pretrained untuk proyeksi query pada mekanisme self-attention, $A_Q \in \mathbb{R}^{r \times d_k}$ adalah matriks down-projection yang memproyeksikan input ke ruang berdimensi rendah r , dan $B_Q \in \mathbb{R}^{d_{\text{model}} \times r}$ adalah matriks up-projection yang mengembalikannya ke dimensi semula. Hasil perkalian $B_Q A_Q$ membentuk matriks low-rank update yang kemudian dijumlahkan ke W_Q sehingga menghasilkan bobot query yang telah disesuaikan, W'_Q , tanpa mengubah bobot pretrained aslinya.

Mekanisme yang sama diterapkan pada proyeksi value, sebagaimana dinyatakan pada Persamaan 3.3:

$$W'_V = W_V + B_V A_V \quad (3.3)$$

Keterangan :

$\sigma(x)$: matriks bobot value dari model pretrained pada mekanisme self-attention

$x \in \mathbb{R}$: matriks bobot value setelah ditambahkan LoRA adapter

$x \in \mathbb{R}$: matriks down-projection LoRA

$x \in \mathbb{R}$: matriks up-projection LoRA

$x \in \mathbb{R}$: rank LoRA

$x \in \mathbb{R}$: matriks update Low Rank

Kunci efisiensi kedua persamaan tersebut terletak pada syarat $r \ll d$, yang berarti rank r jauh lebih kecil dibandingkan dimensi model d . Konsekuensinya,

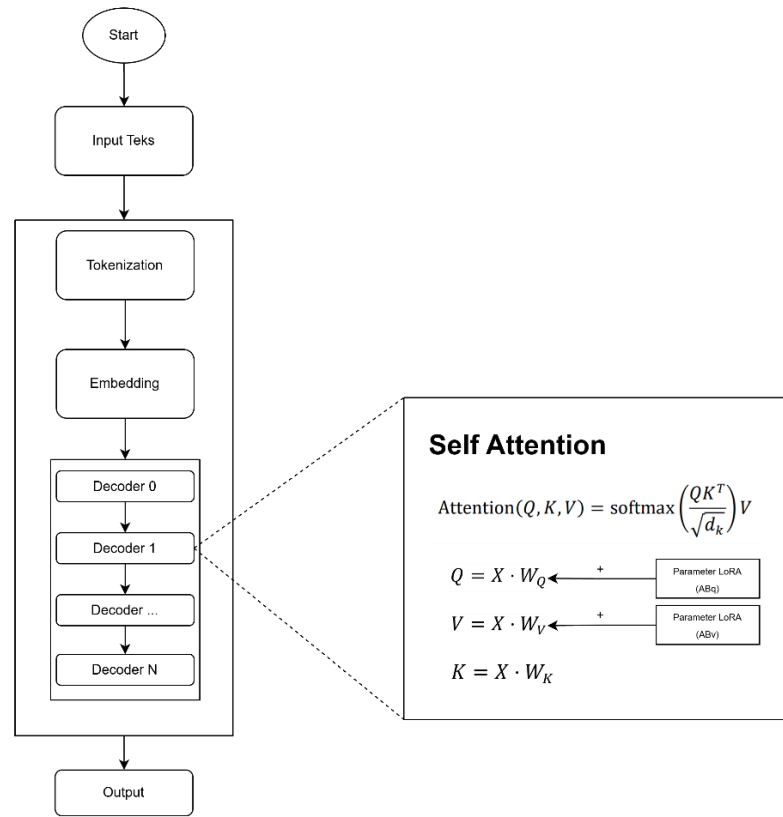
jumlah parameter yang perlu dilatih hanya sebesar $2 \times r \times d$ per lapisan yakni matriks A dan B saja sehingga beban komputasi lebih ringan dibandingkan full fine-tuning, sebagaimana ditunjukkan pada perbandingan parameter di Persamaan 2.3 dan 2.4. Sementara itu, arsitektur decoder-only tetap tidak berubah, sehingga alur attention dari Q, K, V menuju softmax untuk memperoleh nilai atensi tetap berjalan sebagaimana mestinya, namun gaya respons model dapat diarahkan ke domain properti hanya dengan melatih matriks A dan B .

Dengan demikian, Gambar 3.4 dapat dijelaskan sebagai berikut: sisi kiri menunjukkan alur standar decoder-only, di mana input melewati tokenizer, embedding, stack decoder, logits, hingga softmax; sedangkan pada sisi kanan, blok Self-Attention adalah tempat adaptor LoRA ditanamkan tepatnya pada proyeksi linear query dan value melalui penambahan low-rank update AB sebagaimana dirumuskan pada Persamaan 3.12 dan 3.13.

3.3.3 Model TinyLlama-1.1B-Chat-v1.0

Arsitektur yang digunakan oleh model TinyLlama-1.1B-Chat-v1.0 adalah transformer decoder-only seperti yang ditunjukkan pada Gambar 3.4. Aliran datanya berangkat dari input text yang terlebih dahulu diproses oleh tokenizer dahulu menjadi deretan ID token, lalu diubah menjadi bentuk vektor oleh embedding layer.

Vektor dari hasil embedding kemudian masuk ke tumpukan decoder layer seperti Decoder Layer 0 hingga Decoder Layer N. Setiap decoder layer pada dasarnya terdiri dari dua blok utama yaitu self-attention dan feed-forward network. Namun pada penelitian ini akan lebih difokuskan pada bagian self-attention.



Gambar 3. 4 Self Attention

Proses perhitungan matriks Query merupakan bagian awal dari mekanisme self-attention. Matriks Query dihasilkan melalui perkalian antara matriks input X yang merupakan vektor embedding dengan matriks bobot query W_Q . Pada arsitektur ini, parameter LoRA yaitu matriks A dan B ditambahkan pada proyeksi W_Q , sehingga bobot efektif yang digunakan menjadi $W_Q + AB$. Pendekatan ini memungkinkan adaptasi model secara efisien tanpa mengubah seluruh bobot asli TinyLlama-1.1B-Chat-v1.0.

$$Q = X \cdot W_Q \quad (3.4)$$

Keterangan :

Q : Matriks query yang merepresentasikan pencarian konteks dalam suatu pertanyaan

X : Matriks input yang merupakan penggabungan dari vektor embedding

W_Q : Matriks bobot query model

Matriks Key dihitung melalui perkalian matriks input X dengan matriks bobot key W_K . Matriks ini berperan sebagai indeks atau penanda yang akan dicocokkan dengan matriks Query untuk mengukur tingkat kesesuaian antar elemen dalam sequence input.

$$K = X \cdot W_K \quad (3.5)$$

Keterangan :

K : Matriks yang berfungsi sebagai indeks untuk dicocokkan dengan query
 X : Matriks input yang merupakan penggabungan dari vektor embedding
 W_K : Matriks bobot key model

Matriks Value diperoleh melalui perkalian matriks input X dengan matriks bobot value W_V . Sama seperti pada proyeksi Query, parameter LoRA yang digambarkan dengan matriks A dan B, sehingga bobot efektif menjadi $W_V + AB$. Penambahan LoRA pada proyeksi Value memungkinkan model untuk lebih fleksibel dalam mengekstrak informasi aktual yang relevan tanpa meningkatkan beban komputasi secara signifikan.

$$V = X \cdot W_V \quad (3.6)$$

Keterangan :

V : informasi aktual yang akan diambil setelah skor attention dihitung
 X : Matriks input yang merupakan penggabungan dari vektor embedding
 W_V : matriks bobot Value model

Rumus self-attention di atas menggambarkan mekanisme inti dari decoder layer. Skor kesesuaian antara Query dan Key dihitung melalui perkalian dot-product, kemudian diskalakan dengan faktor $\sqrt{d_k}$ untuk menjaga stabilitas gradien, dan dinormalisasi menggunakan fungsi softmax. Hasil normalisasi tersebut dikalikan dengan matriks Value untuk menghasilkan representasi kontekstual yang

telah diberi bobot perhatian. Mekanisme self-attention ini sangat berpengaruh terhadap performa keseluruhan model karena memungkinkan setiap token untuk “memperhatikan” seluruh token lain dalam sequence, sehingga model mampu menangkap ketergantungan jarak jauh dan konteks yang kompleks kemampuan utama yang membuat arsitektur transformer decoder-only seperti TinyLlama bisa melakukan tugas generasi teks.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.7)$$

Keterangan:

Q : Matriks query yang digunakan untuk mencari informasi yang relevan

K : Matriks key yang merepresentasikan karakteristik atau penanda dari setiap elemen input.

V : Matriks value berisi informasi aktual yang akan digabungkan berdasarkan bobot perhatian.

T : Transpose Matriks.

d_k : Dimensi vektor *key*, sebagai faktor penskalaan untuk menjaga stabilitas nilai hasil perkalian.

$\text{softmax}()$: Fungsi normalisasi yang mengubah skor kesesuaian menjadi bobot probabilitas.

Setelah melewati seluruh tumpukan decoder layer (Decoder 0 hingga Decoder N), keluaran akhir berupa representasi vektor kontekstual akan diproses lebih lanjut oleh language modeling head. Head ini menghasilkan logits yang merepresentasikan distribusi probabilitas prediksi token selanjutnya. Logits dirubah ke bentuk nilai probabilitas dengan menggunakan softmax lalu selanjutnya dibandingkan dengan token target yang sebenarnya dari dataset pelatihan melalui fungsi Cross Entropy Loss. Nilai loss yang dihasilkan kemudian menjadi dasar proses Backpropagation untuk memperbarui parameter LoRA pada layer attention, sesuai dengan alur desain sistem pelatihan yang telah diuraikan sebelumnya.

3.3.4 Cross Entropy Loss

Setelah model menghasilkan probabilitas melalui fungsi softmax, diperlukan suatu cara untuk mengukur seberapa jauh prediksi model dari jawaban yang sebenarnya. Fungsi yang digunakan untuk keperluan ini adalah cross-entropy loss. Fungsi ini bekerja dengan cara mengambil nilai logaritma dari probabilitas yang diberikan model terhadap kelas yang benar, kemudian memberikan tanda negatif agar nilai yang dihasilkan selalu positif.

$$L = -\log(p_y) \quad (3.8)$$

Keterangan:

L : Nilai *loss* (fungsi kerugian) yang digunakan untuk mengukur kesalahan prediksi model
 p_y : Probabilitas prediksi model terhadap kelas yang benar
 $\log$: Fungsi logaritma

Semakin tinggi keyakinan model terhadap kelas yang benar, maka nilai loss akan semakin kecil dan mendekati nol. Sebaliknya, apabila model memberikan probabilitas yang rendah pada kelas yang benar, nilai loss akan menjadi besar, yang menjadi sinyal bagi model bahwa prediksinya masih jauh dari target.

3.3.5 Backpropagation

Setelah nilai loss diperoleh, langkah berikutnya adalah mempropagasikan kesalahan tersebut mundur melalui jaringan agar parameter yang dapat dilatih dapat diperbarui. Proses ini dikenal sebagai backpropagation. Dalam konteks LoRA, langkah pertama backpropagation adalah menghitung gradien loss terhadap logits. Gradien ini menunjukkan seberapa besar dan ke arah mana setiap logit perlu diubah agar loss berkurang. Berkat sifat matematis dari kombinasi softmax dan cross-entropy, gradien terhadap logits memiliki bentuk yang sangat sederhana, yaitu

selisih antara probabilitas prediksi model dengan label sebenarnya dalam format one-hot.

$$\frac{\partial L}{\partial z_i} = p_i - y_i \quad (3.9)$$

Keterangan:

L : Fungsi loss (kerugian model)
 z_i : Logit untuk kelas ke- i
 p_i : Probabilitas hasil softmax untuk kelas ke- i
 y_i : Label one-hot untuk kelas ke- i
 $\frac{\partial L}{\partial z_i}$: Gradien loss terhadap logit

Ketika gradien terhadap logits diperoleh, nilai tersebut selanjutnya dipropagasikan lebih jauh ke belakang hingga mencapai parameter LoRA. Karena bobot pretrained dalam keadaan dibekukan, gradien tidak mengalir ke dalamnya dan hanya mengalir ke parameter LoRA yang merupakan satu-satunya parameter yang dapat dilatih. Pembaruan parameter LoRA dilakukan menggunakan metode optimasi AdamW, yaitu varian dari optimizer Adam yang menggabungkan mekanisme momentum dan adaptive learning rate sekaligus menerapkan weight decay secara langsung pada bobot untuk mencegah overfitting, sehingga proses konvergensi selama fine-tuning diharapkan dapat berlangsung lebih terkontrol.

Dengan diperbaruinya seluruh parameter LoRA secara bersamaan, matriks update ΔW pada iterasi berikutnya akan menghasilkan koreksi yang lebih tepat terhadap bobot query, sehingga model secara bertahap belajar menyesuaikan diri dengan domain data fine-tuning. Satu siklus yang mencakup forward pass, perhitungan loss, hingga pembaruan seluruh parameter LoRA ini disebut satu iterasi training. Proses tersebut diulang untuk seluruh data training hingga nilai loss

konvergen, yang menandakan bahwa model telah berhasil beradaptasi dengan domain baru melalui parameter LoRA yang telah terlatih.

3.3.6 LoRA Parameter

Setelah seluruh siklus pelatihan selesai dijalankan dan nilai loss telah konvergen, hasil akhir dari proses fine-tuning berbasis LoRA ini bukan berupa pembaruan pada bobot model TinyLLaMA secara keseluruhan, melainkan dua matriks ringkas yang menjadi inti dari adaptor LoRA, yaitu matriks A dan matriks B . Kedua matriks inilah yang selama proses pelatihan secara iteratif diperbarui melalui backpropagation, sementara seluruh bobot pretrained model tetap dibekukan dan tidak mengalami perubahan apapun.

Matriks $A \in \mathbb{R}^{r \times d}$ dan matriks $B \in \mathbb{R}^{d \times r}$ bersama-sama membentuk matriks low-rank update $\Delta W = AB$, yang merepresentasikan penyesuaian domain yang dipelajari model selama fine-tuning. Perkalian kedua matriks ini, yakni AB , menghasilkan matriks berukuran penuh $\mathbb{R}^{d \times d}$ yang merepresentasikan pola domain properti yang dipelajari dari data pelatihan, namun dengan jumlah parameter yang jauh lebih kecil dibandingkan bobot asli model. Sebagai ilustrasi, apabila dimensi model $d = 2048$ dan rank yang digunakan $r = 8$, maka total parameter yang dilatih hanya sebesar $2 \times 8 \times 2048 = 32.768$ per lapisan, dibandingkan $2048 \times 2048 = 4.194.304$ parameter apabila dilakukan full fine-tuning.

Adaptor LoRA yang dihasilkan, yaitu pasangan matriks $A_Q B_Q$ untuk proyeksi query dan $A_V B_V$ untuk proyeksi value, kemudian disimpan sebagai parameter tambahan yang terpisah dari bobot utama model. Pada saat inferensi, adaptor ini digabungkan kembali ke dalam bobot pretrained melalui operasi

penjumlahan $W' = W + BA$ sebagaimana dirumuskan pada Persamaan 3.12 dan 3.13, sehingga model mampu menghasilkan respons yang telah diadaptasi ke domain properti tanpa memerlukan beban memori maupun komputasi yang besar. Dengan demikian, hasil akhir dari keseluruhan sistem pelatihan ini adalah sepasang adaptor LoRA yang ringan, portabel, dan dapat digunakan ulang, yang menjadi jembatan antara kapabilitas umum model TinyLLaMA pretrained dengan kebutuhan spesifik domain pembelian properti, simulasi KPR, legalitas dokumen, biaya, dan pajak.

3.4 Skenario Uji coba

Pada penelitian ini, skenario uji disusun untuk membandingkan performa model berdasarkan variasi konfigurasi hyperparameter LoRA yang diterapkan selama proses fine-tuning. Pengujian dilakukan pada tiga skenario, yaitu Skenario 1, Skenario 2, serta Skenario 3 sebagai model hasil fine-tuning menggunakan pendekatan Low-Rank Adaptation (LoRA) pada model TinyLLaMA-1.1B-Chat-v1.0. Setiap skenario menggunakan dataset yang sama, namun dibedakan berdasarkan konfigurasi pelatihan seperti rank dan alpha. Setelah proses pelatihan dilakukan, setiap model dievaluasi berdasarkan hasil pelatihannya melalui pengujian kualitas respons menggunakan metrik BLEU-4. Rangkaian pengujian ini bertujuan untuk membandingkan performa antar model, menentukan model yang paling optimal, dan selanjutnya melakukan pengujian implementasi model terpilih pada sistem chatbot konsultan properti untuk dibandingkan performanya dengan agen property.

3.4.1 Konfigurasi Pelatihan Model

Konfigurasi skenario dibagi menjadi tiga skenario uji menggunakan pendekatan Low-Rank Adaptation (LoRA), sebagaimana ditunjukkan pada Tabel 3.5. Perbedaan antar skenario terletak pada pengaturan hyperparameter utama, yaitu LoRA Rank (r) dan LoRA Alpha (α), sementara hyperparameter lainnya dikontrol secara konsisten di seluruh skenario.

Tabel 3. 5 Skenario Uji

Skenario Uji	LoRA Rank (r)	LoRA Alpha	LoRA Dropout	Learning Rate	Epoch
Skenario 1	8	16	0	2×10^{-4}	12
Skenario 2	16	32	0	2×10^{-4}	12
Skenario 3	32	64	0	2×10^{-4}	12

Pada penelitian Hu et al. (2021) dalam ablation study menunjukkan bahwa performa LoRA tidak meningkat signifikan ketika rank dinaikkan melampaui $r = 4$ hingga $r = 8$, bahkan $r = 64$ pun hasilnya tetap kompetitif tanpa perbaikan berarti. Hal ini dipertegas oleh pernyataan mereka bahwa meningkatkan nilai rank tidak selalu membawa perubahan yang signifikan, serta didukung hasil yang menunjukkan validation loss terendah berada pada $r = 16$ dan tidak membaik di atasnya. Biderman et al. (2024) pada turut mengonfirmasi bahwa $r = 16$ sudah menghasilkan perbaikan nyata pada fine-tuning LoRA, sementara kenaikan rank lebih lanjut hanya memberikan peningkatan moderat. Berdasarkan temuan-temuan tersebut, penelitian ini menetapkan nilai rank $r = 8$, $r = 16$, dan $r = 32$ sebagai representasi tiga kelas kapasitas adaptasi dalam rentang yang terbukti efektif secara empiris.

Nilai LoRA Alpha pada setiap skenario ditetapkan sebesar dua kali nilai rank ($\alpha = 2r$), menghasilkan faktor skala konstan sebesar 2 di seluruh skenario. Penetapan ini mengikuti rekomendasi eksplisit Biderman et al. (2024) secara empiris mereka membuktikan bahwa $\alpha = 2r$ menghasilkan akurasi tertinggi dibanding konfigurasi α lainnya pada eksperimen code dan math IFT. Dengan $\alpha = 2r$, perbedaan performa yang terukur antar skenario dapat dikaitkan secara langsung pada variasi rank, bukan pada interaksi antara rank dan alpha.

LoRA Dropout ditetapkan pada nilai 0 di seluruh skenario. Keputusan ini didasarkan pada ukuran dataset yang terbatas (~3000 sampel), di mana penerapan dropout berisiko menyebabkan underfitting karena model tidak memiliki cukup data untuk mengompensasi informasi yang dihilangkan secara acak selama pelatihan. Hal ini sejalan dengan praktik pada Hu et al. (2021) yang dalam seluruh konfigurasi GPT-2 mereka tidak menerapkan dropout tambahan pada modul LoRA.

Learning rate sebesar 2×10^{-4} dipilih berdasarkan rentang yang direkomendasikan Biderman et al. (2024), yang menyatakan bahwa learning rate optimal untuk LoRA pada skenario instruction fine-tuning berada antara 5×10^{-4} hingga 5×10^{-4} , dengan nilai terbaik yang ditemukan secara konsisten pada kisaran 1×10^{-4} hingga 2×10^{-4} untuk model berukuran kecil hingga menengah. Nilai 2×10^{-4} dipilih sebagai titik tengah yang stabil dalam rentang tersebut.

Jumlah epoch sebesar 12 ditetapkan dengan mempertimbangkan temuan Biderman et al. (2024) yang menunjukkan bahwa model LoRA dengan $r = 16$ pada dataset instruction fine-tuning mencapai performa puncak pada epoch ke-4 hingga ke-8, sementara konvergensi penuh membutuhkan epoch lebih banyak dibanding

full fine-tuning karena kapasitas parameter yang lebih terbatas. Dengan dataset yang jauh lebih kecil (~3000 sampel dibanding ~100K pada Biderman et al.), model membutuhkan lebih banyak iterasi untuk mengekstrak pola yang memadai dari data, sehingga 12 epoch ditetapkan sebagai jumlah yang memberikan ruang konvergensi yang cukup sekaligus masih dalam batas yang dapat dipantau untuk menghindari overfitting.

3.4.2 Uji Hasil Pelatihan Model

Uji hasil pelatihan model pada penelitian ini dilakukan melalui dua aspek utama, yaitu pengujian kualitas respons model. Pengujian kualitas respons digunakan untuk menilai sejauh mana model mampu menghasilkan jawaban yang relevan, sesuai konteks, dan mendekati jawaban referensi pada domain properti. Pada tahap ini, evaluasi dilakukan terhadap seluruh skenario model dengan menggunakan metrik BLEU-4 untuk mengukur tingkat kesesuaian tekstual keluaran model terhadap data referensi pada domain properti. Melalui pengujian ini, hasil pelatihan tidak hanya dilihat dari pola konvergensi training loss, tetapi juga dari sejauh mana keluaran model mendekati jawaban referensi yang telah ditetapkan.

3.4.3 Pemilihan Model Terbaik

Model terbaik ditentukan berdasarkan kombinasi dua aspek utama, yaitu skor BLEU-4 sebagai indikator kesesuaian tekstual keluaran model terhadap data referensi, serta pola konvergensi training loss sebagai cerminan stabilitas proses pelatihan. Kedua aspek ini dievaluasi secara bersamaan agar pemilihan model tidak

semata-mata bergantung pada satu dimensi pengukuran, melainkan mempertimbangkan konsistensi pembelajaran model selama proses fine-tuning sekaligus kualitas keluaran yang dihasilkan.

Meskipun skor BLEU-4 tidak secara langsung menjamin kualitas jawaban secara semantik, metrik ini digunakan sebagai acuan kuantitatif yang objektif dan konsisten untuk membandingkan ketiga skenario dalam kondisi dataset dan konfigurasi pelatihan yang identik. Dengan pendekatan ini, model yang menunjukkan skor BLEU-4 tertinggi diantara skenario lainnya sekaligus pola penurunan loss yang stabil dipilih sebagai model terbaik untuk selanjutnya digunakan pada tahap uji implementasi dalam sistem chatbot konsultan properti.

3.4.4 Uji Implementasi Model

Uji implementasi model pada penelitian ini menggunakan kuesioner untuk membandingkan performa chatbot berbasis TinyLLaMA yang telah di fine-tune menggunakan LoRA dengan agen manusia dalam konteks layanan properti. Pengujian dilakukan menggunakan instrumen berupa kuesioner pengguna yang dirancang untuk menangkap pengalaman langsung responden setelah berinteraksi dengan kedua sistem. Interaksi tersebut difokuskan pada tujuh topik utama, yaitu KPR dan pembiayaan properti, legalitas properti, biaya dan pajak, spesifikasi properti, proses transaksi, konsultasi keputusan, serta risiko dan disclaimer. Melalui skenario ini, penilaian implementasi model tidak hanya dilakukan dari sisi teknis, tetapi juga dari sisi pengalaman pengguna dalam menerima layanan konsultasi properti.

Setiap responden diminta untuk melakukan percakapan dengan chatbot dan agen manusia pada salah satu topik yang telah ditentukan, kemudian mengisi kuesioner berdasarkan pengalaman interaksi tersebut. Kuesioner disusun untuk mengukur satu aspek saja, kualitas jawaban yang dihasilkan oleh model. Kualitas jawaban dievaluasi melalui tingkat kelengkapan, kemudahan pemahaman, dan kepuasan pengguna terhadap jawaban yang diberikan. Selain membandingkan tingkat kepuasan terhadap chatbot dan agen manusia, kuesioner ini juga memuat pertanyaan perbandingan untuk mengetahui sistem mana yang dinilai lebih cepat atau lebih mudah digunakan oleh responden.

Sebagai pelengkap, kuesioner juga menyediakan bagian umpan balik terbuka agar responden dapat menyampaikan pendapat, kritik, atau saran terhadap chatbot yang diuji. Struktur pertanyaan dalam kuesioner disusun dalam bentuk skala penilaian untuk memudahkan proses analisis, sebagaimana disajikan pada Tabel 3.8. Melalui uji implementasi ini, peneliti dapat memperoleh data yang terstruktur mengenai persepsi pengguna terhadap kecepatan dan kualitas layanan dari kedua sistem, sekaligus mengidentifikasi aspek-aspek yang masih perlu diperbaiki. Dengan demikian, evaluasi implementasi model dilakukan secara lebih komprehensif karena didasarkan pada pengalaman pengguna secara langsung dalam konteks layanan properti.

3.5 Evaluasi BLEU

Evaluasi dilakukan untuk menilai kemampuan model hasil fine-tuning dalam menghasilkan teks yang relevan dan sesuai konteks. Metrik yang digunakan adalah BLEU-4, yang populer untuk menilai kualitas keluaran model NLP.

BLEU-4 atau Bilingual Evaluation Understudy menilai kesamaan n-gram dengan fokus pada precision untuk menghitung sejauh mana kata atau frasa pada prediksi muncul dalam reference. BLEU-4 juga menerapkan brevity penalty untuk menghindari prediksi yang terlalu pendek.

Dalam praktiknya, teks keluaran model dibandingkan dengan gold standard menggunakan pustaka datasets dan evaluate dari Hugging Face. Berikut dibawah inii merupakan contoh penghitungan sederhana:

Tabel 3. 6 Contoh Token Perhitungan

Kalimat	Token (setelah normalisasi)	Jumlah
Referensi (R)	simulasi, kpr, untuk, rumah, seharga, tiga, ratus, juta, dengan, uang, muka, dua, puluh, persen, selama, dua, puluh, tahun, akan, menghasilkan, cicilan, sekitar, dua, juta, per, bulan	26 token
Candidate (C)	untuk, kpr, rumah, harga, tiga, ratus, juta, bisa, dicicil, sekitar, dua, juta, rupiah, tiap, bulan	15 token

Tabel 3. 7 Contoh Perhitungan BLEU-4

Komponen	Detail / Rumus	Nilai
Brevity Penalty (BP)	$ C =15, R =26 \rightarrow c < r \rightarrow BP = \exp(1 - 26/15) = \exp(-0.733)$	≈ 0.481
Precision 1-gram (p_1)	9 match / 15 token prediksi	≈ 0.600
Precision 2-gram (p_2)	3 bigram match / 14 bigram prediksi	≈ 0.214
Precision 3-gram (p_3)	"tiga ratus juta", "sekitar dua juta" $\rightarrow$ 2 match / 13	≈ 0.154
Precision 4-gram (p_4)	"tiga ratus juta bisa" $\neq$ ref $\rightarrow$ 0 match / 12	0.000
BLEU-4	$BP \times \exp(\frac{1}{4} \times [\ln p_1 + \ln p_2 + \ln p_3 + \ln p_4])$ $\rightarrow$ karena $p_4 = 0$, nilai BLEU-4 = 0 untuk pasangan ini	≈ 0 (sangat rendah)

Nilai BLEU-4 = 0 pada contoh tabel diatas ini terjadi karena tidak ada 4-gram yang cocok persis antara prediksi dan referensi. Dalam evaluasi sesungguhnya pada BAB 4, skor dihitung secara agregat dari seluruh pasangan data sehingga menghasilkan nilai rata-rata yang lebih representatif.

Penggunaan metrik BLEU-4 berfokus pada pengukuran precision melalui perbandingan n-gram antara teks hasil model dan teks referensi, sehingga kualitas model dapat dievaluasi berdasarkan tingkat kesesuaian output yang dihasilkan.

3.6 Evaluasi Pengguna

Dalam penelitian ini, untuk membandingkan kualitas jawaban antara chatbot berbasis TinyLLaMA yang telah di fine-tune dengan LoRA dan agen manusia dalam konteks layanan properti, digunakan instrumen berupa kuesioner pengguna. Kuesioner ini dirancang untuk menangkap pengalaman langsung pengguna setelah berinteraksi dengan kedua sistem pada tujuh topik utama, yaitu KPR dan pembiayaan properti, legalitas properti, biaya dan pajak, spesifikasi properti, proses transaksi, konsultasi keputusan, serta risiko dan disclaimer.

Setiap responden akan melakukan percakapan dengan chatbot dan agen manusia pada salah satu topik tersebut, kemudian mengisi kuesioner yang mengukur satu aspek saja, yaitu kualitas jawaban yang dihasilkan oleh model. Kualitas jawaban dievaluasi melalui akurasi, kelengkapan, dan kemudahan pemahaman. Selain itu, kuesioner juga mengukur tingkat kepuasan serta memberikan ruang bagi responden untuk menyampaikan umpan balik kualitatif. Struktur kuesioner disajikan pada Tabel berikut.

Tabel 3. 8 Contoh Quesioner

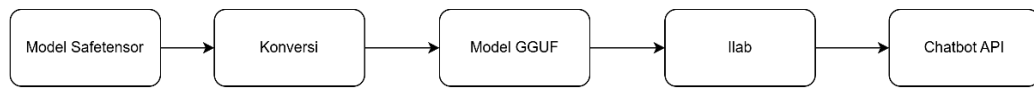
Bagian	Pertanyaan	Pilihan Jawaban
Kualitas Jawaban	Seberapa lengkap jawaban yang diberikan?	1. Sangat Lengkap 2. Lengkap 3. Cukup Lengkap 4. Tidak Lengkap 5. Sangat Tidak Lengkap
	Apakah jawaban mudah dipahami?	1. Sangat Mudah Dipahami 2. Mudah Dipahami

		3. Cukup Mudah Dipahami 4. Sulit Dipahami 5. Sangat Sulit Dipahami
Kepuasan	Berikan tingkat kepuasan anda terhadap chatbot	1. Sangat Puas 2. Puas
	Berikan tingkat kepuasan anda terhadap agen manusia	3. Cukup Puas 4. Tidak Puas 5. Sangat Tidak Puas

Melalui kuesioner ini peneliti mengumpulkan jawaban yang terstruktur dari pengguna untuk membandingkan performa chatbot dan agen manusia. Penilaian dilakukan berdasarkan persepsi pengguna terhadap kualitas layanan, serta untuk mengidentifikasi aspek-aspek yang perlu diperbaiki pada kedua sistem. Pendekatan ini memastikan evaluasi berbasis pengalaman pengguna yang komprehensif dan relevan dengan konteks layanan properti.

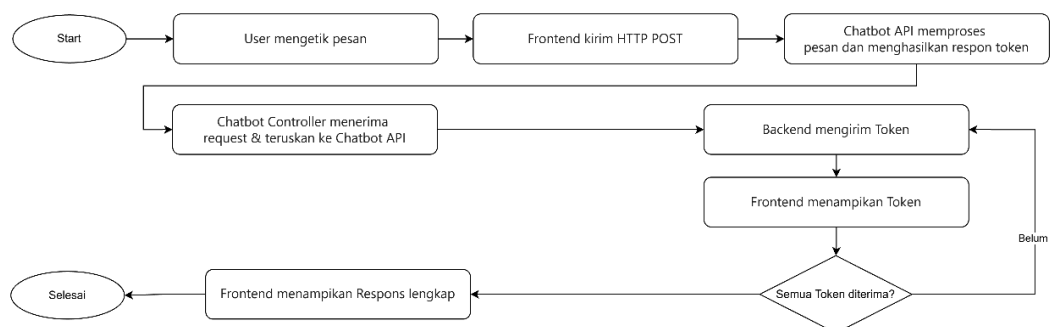
3.7 Implementasi Model Chatbot

Pada bagian ini, akan dibahas mengenai bagaimana model chatbot diimplementasikan dalam sistem menggunakan framework Laravel dan tools tambahan seperti ilab menjalankan chatbot. Tujuan utama dari implementasi ini adalah untuk memungkinkan chatbot memberikan respons otomatis kepada pengguna berdasarkan permintaan yang diterima. Proses dimulai dengan mengkonversi model hasil training ke format GGUF dan memuat model ke dalam server yang disediakan oleh ilab, lalu model siap untuk menerima dan memproses permintaan API melalui Chatbot API. Berikut adalah penjelasan detail mengenai implementasi model ini dan komponen-komponen yang terlibat dalam sistem.



Gambar 3. 5 Alur Menjalankan Chatbot

Gambar di atas menggambarkan alur dalam menjalankan model chatbot yang melibatkan lima komponen utama. Proses dimulai dengan Model Safetensor, yang merujuk pada model yang telah dilatih dan disimpan dalam format Safetensor. Model ini berfungsi sebagai dasar bagi chatbot untuk melakukan inferensi. Setelah itu, model ini diproses lebih lanjut melalui proses konversi dengan menggunakan library llama-cpp untuk memastikan kompatibilitas dengan platform yang digunakan. Kemudian, model yang sudah melalui proses konversi disimpan dalam Model GGUF. Selanjutnya, llab bertindak sebagai server yang menyiapkan model untuk di hosting, agar siap menerima permintaan dari aplikasi luar. Akhirnya, Chatbot API menjadi antarmuka yang memungkinkan pengguna untuk berinteraksi dengan model chatbot, mengirimkan permintaan, dan menerima respons yang dihasilkan oleh model.



Gambar 3. 6 Flowchart Alur Komunikasi Chatbot pada Sistem

Pada bagian ini, akan dijelaskan alur komunikasi yang terjadi antara pengguna, antarmuka pengguna atau UI, backend Laravel (ChatbotController), dan

API chatbot eksternal. Sequence diagram yang ditampilkan menggambarkan tahapan interaksi yang berlangsung, mulai dari pengguna mengirimkan pesan hingga menerima respons dari chatbot. Diagram ini memberikan gambaran tentang bagaimana sistem beroperasi dalam menyampaikan dan memproses pesan, serta cara data dikirimkan dan diterima antar berbagai komponen sistem. Berikut penjelasan detail mengenai langkah-langkah yang terjadi dalam proses ini, yang dapat dijadikan bagian dari penelitian ini:

1. User mengirimkan pesan dengan mengetikkan input pada kolom yang tersedia pada Frontend.
2. Frontend menampilkan pesan dari pengguna pada chat box sebagai pesan pengguna dan kemudian mengirimkan pesan tersebut melalui permintaan HTTP POST ke endpoint /chat-stream yang ada pada ChatbotController di backend Laravel.
3. Backend menerima permintaan POST yang berisi pesan dari pengguna dan meneruskannya ke Chatbot API, dengan menyertakan pesan yang diterima dalam bentuk request.
4. Chatbot API memproses pesan yang diterima dan menghasilkan respons berupa token atau konten respons yang kemudian dikirimkan kembali ke system backend.
5. Backend kemudian mengirimkan token respons yang diterima dari Chatbot API ke Frontend melalui mekanisme server-sent events, yang memungkinkan pengiriman data secara bertahap tanpa perlu penyegaran ulang pada sisi klien.

6. Frontend kemudian menampilkan respons token chatbot secara dinamis pada chat box, sehingga pengguna dapat melihat respons yang diterima secara langsung.
7. Setelah seluruh token diterima dan ditampilkan, Frontend mengirimkan respons lengkap kembali ke Backend dan menampilkan pesan akhir yang berupa respons chatbot kepada pengguna dalam tampilan chat.

Diagram ini menggambarkan alur komunikasi yang menghubungkan antarmuka pengguna dengan backend dan API chatbot eksternal, dengan memanfaatkan SSE untuk streaming respons secara real-time.

3.8 Contoh Perhitungan Manual Fine-Tuning LoRA

Untuk memperjelas mekanisme fine-tuning menggunakan Low-Rank Adaptation, pada bagian ini disajikan simulasi perhitungan manual dengan dimensi yang disederhanakan. Contoh ini menggunakan data percakapan domain properti dengan intent `kpr_simulasi`, berupa pertanyaan pengguna:

"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?"

Untuk keluaran target yang diharapkan adalah respons berisi simulasi KPR lengkap mencakup perhitungan DP, plafon pinjaman, estimasi cicilan, hingga rasio terhadap gaji. Dalam contoh ini digunakan dimensi model $d_{model} = 4$, dimensi proyeksi query $d_k = 4$, serta rank LoRA $r = 2$. Panjang urutan diasumsikan $L = 1$ agar perhitungan dapat dilakukan secara bertahap dan sistematis. Ukuran kosakata ditetapkan sebanyak 3 untuk keperluan ilustrasi softmax dan cross-entropy.

3.8.1 Representasi Input (Embedding)

Setelah kalimat instruksi melewati proses tokenisasi, salah satu token kunci misalkan token yang merepresentasikan kata "cicilan" menghasilkan vektor embedding:

$$x = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}$$

Vektor tersebut berada pada ruang berdimensi empat, sehingga $x \in \mathbb{R}^4$. Nilai-nilai ini merupakan representasi numerik fitur token yang akan diproyeksikan ke ruang query untuk memproses konteks pertanyaan KPR dari pengguna.

3.8.2 Bobot Pretrained Query (Dibekukan)

Pada metode LoRA, bobot pretrained tidak diperbarui selama proses fine-tuning. Misalkan bobot pretrained untuk proyeksi query berbentuk matriks identitas:

$$W_q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Karena matriks ini identitas, apabila tidak digunakan LoRA maka hasil proyeksi query akan identik dengan vektor input x . Namun dalam pendekatan LoRA, bobot efektif dihitung sebagaimana dirumuskan pada Persamaan 3.12, sehingga model dapat mempelajari pola respons domain KPR hanya melalui matriks A_Q dan B_Q tanpa mengubah W_Q .

3.8.3 Parameter LoRA

Dengan rank $r = 2$, maka dimensi parameter LoRA untuk proyeksi query adalah:

$$B \in \mathbb{R}^{2 \times 4} \text{ dan } A \in \mathbb{R}^{4 \times 2}$$

Misalkan diberikan:

$$B = \begin{bmatrix} 0.1 & 0.2 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.2 & 0.1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0.3 & 0.1 \\ 0.2 & 0.4 \\ 0.5 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}$$

Matriks B_Q berfungsi sebagai down-projection dari dimensi 4 ke 2, menangkap fitur-fitur penting dari token pertanyaan KPR. Sedangkan A_Q mengembalikannya dari dimensi 2 ke 4, menghasilkan koreksi bobot yang mengarahkan model untuk merespons dalam konteks simulasi KPR.

3.8.4 Perhitungan Update Low-Rank (BA)

Update parameter dihitung melalui perkalian matriks:

$$BA = B \cdot A$$

Hasil perkalian tersebut adalah:

$$BA = \begin{bmatrix} 0.03 & 0.07 & 0.05 & 0.01 \\ 0.02 & 0.08 & 0.10 & 0.04 \\ 0.05 & 0.12 & 0.09 & 0.02 \\ 0.01 & 0.05 & 0.07 & 0.03 \end{bmatrix}$$

Matriks ini berdimensi 4×4 , sama seperti W_q , sehingga dapat dijumlahkan langsung.

3.8.5 Pembentukan Bobot Efektif

Bobot query yang digunakan dalam forward pass menjadi:

$$W'_q = W_q + BA$$

Sehingga diperoleh:

$$W'_q = \begin{bmatrix} 1.03 & 0.07 & 0.05 & 0.01 \\ 0.02 & 1.08 & 0.10 & 0.04 \\ 0.05 & 0.12 & 1.09 & 0.02 \\ 0.01 & 0.05 & 0.07 & 1.03 \end{bmatrix}$$

Terlihat bahwa setiap elemen diagonal mengalami sedikit peningkatan, sementara elemen non-diagonal mendapatkan kontribusi koreksi kecil dari matriks low-rank.

3.8.6 Forward Pass: Perhitungan Query

Selanjutnya dihitung vektor query:

$$Q = W'_q x$$

Dengan:

$$x = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}$$

Perhitungan dilakukan per baris:

Baris 1:

$$(1.03)(1) + (0.07)(0) + (0.05)(2) + (0.01)(1) = 1.03 + 0 + 0.10 + 0.01 = 1.14$$

Baris 2:

$$(0.02)(1) + (1.08)(0) + (0.10)(2) + (0.04)(1) = 0.02 + 0 + 0.20 + 0.04 = 0.26$$

Baris 3:

$$(0.05)(1) + (0.12)(0) + (1.09)(2) + (0.02)(1) = 0.05 + 0 + 2.18 + 0.02 = 2.25$$

Baris 4:

$$(0.01)(1) + (0.05)(0) + (0.07)(2) + (1.03)(1) = 0.01 + 0 + 0.14 + 1.03 = 1.18$$

Sehingga diperoleh:

$$Q = \begin{bmatrix} 1.14 \\ 0.26 \\ 2.25 \\ 1.18 \end{bmatrix}$$

Vektor query ini merepresentasikan "pencarian konteks" dari token "cicilan" setelah bobot disesuaikan oleh LoRA, sehingga mekanisme attention dapat lebih fokus menangkap hubungan antara besaran gaji, harga rumah, dan kelayakan cicilan dalam konteks KPR.

3.8.7 Perhitungan Logits dan Softmax

Misalkan setelah seluruh decoder layer diproses, model menghasilkan logits untuk tiga kandidat token berikutnya yang merepresentasikan kemungkinan awal respons model:

Tabel 3. 9 Permisalan Nilai Logits

Token	Representasi	Logit
Token 1	"Maaf,..."	$z_1 = 1.5$
Token 2	"Baik,..."	$z_2 = 0.5$
Token 3	"Tentu,..."	$z_3 = 2.0$

Nilai eksponensialnya adalah:

$$e^{1.5} = 4.48$$

$$e^{0.5} = 1.65$$

$$e^{2.0} = 7.39$$

Total:

$$4.48 + 1.65 + 7.39 = 13.52$$

Probabilitas softmax:

$$p_1 = 0.331$$

$$p_2 = 0.122$$

$$p_3 = 0.547$$

Pada tahap ini model cenderung memilih token ketiga ("Tentu,..."), padahal token target yang benar berdasarkan data pelatihan adalah token kedua ("Baik,...") yakni awal dari respons target "Baik, saya bantu dengan simulasi sederhana...".

3.8.8 Perhitungan Cross-Entropy Loss

Karena token target yang benar adalah token ke-2 ("Baik,..."), maka loss dihitung berdasarkan probabilitas yang diberikan model pada token tersebut:

$$Loss = -\log(0.122) = 2.10$$

Nilai loss sebesar 2.10 ini menunjukkan bahwa model masih memberikan probabilitas yang rendah pada token yang benar. Artinya, pada iterasi ini model belum mampu memprediksi bahwa respons yang tepat untuk pertanyaan simulasi KPR adalah yang diawali dengan kalimat konfirmatif "Baik, saya bantu..." beserta rincian DP, plafon, dan estimasi cicilan.

3.8.9 Perhitungan Gradien

Dengan target 'one-hot' $y = [0,1,0]$ dimana token ke-2 adalah yang benar, gradien terhadap logits dihitung sebagai selisih probabilitas prediksi dengan label:

$$\frac{\partial L}{\partial z_1} = 0.331$$

$$\frac{\partial L}{\partial z_2} = -0.878$$

$$\frac{\partial L}{\partial z_3} = 0.547$$

Gradien negatif pada z_2 menunjukkan bahwa bobot yang mendukung token "Baik,..." perlu diperkuat, sementara gradien positif pada z_1 dan z_3 menunjukkan

bahwa bobot yang mengarah ke token lain perlu dikurangi. Karena W_Q dibekukan, gradien ini hanya mengalir ke parameter LoRA A_Q dan B_Q .

3.8.10 Update Parameter LoRA

Dengan learning rate:

$$\eta = 0.01$$

Dan gradien terhadap A :

$$\frac{\partial L}{\partial A} = \begin{bmatrix} 0.2 & 0.1 & 0.3 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.2 \end{bmatrix}$$

Maka pembaruan dilakukan sebagai:

$$A_{Q \text{ baru}} = A_Q - \eta \cdot \frac{\partial L}{\partial A_Q}$$

Contoh elemen pertama:

$$0.1 - 0.01(0.2) = 0.098$$

Proses yang sama diterapkan pada seluruh elemen A_Q dan B_Q . Setelah pembaruan ini, matriks low-rank update $B_Q A_Q$ pada iterasi berikutnya akan menghasilkan bobot efektif W'_Q yang sedikit lebih optimal, sehingga probabilitas token "Baik,..." meningkat dan model secara bertahap belajar menghasilkan respons simulasi KPR yang sesuai dengan pola pada data pelatihan mencakup komponen DP sekitar Rp48,75 juta, plafon pinjaman Rp276 juta, estimasi cicilan Rp2,4–2,7 juta per bulan, hingga rasio cicilan terhadap gaji sebesar 28–31%.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Gambaran Umum Penelitian

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi sistem chatbot konsultan properti berbasis Large Language Model dengan memanfaatkan arsitektur TinyLLaMA yang telah di-fine-tuning menggunakan metode Low-Rank Adaptation atau dapat disingkat dengan LoRA. Fokus utama penelitian adalah menghasilkan model yang mampu memberikan respons konsultatif, edukatif, dan kontekstual dalam domain properti, khususnya terkait pembelian rumah dan pengambilan keputusan finansial. Proses penelitian diawali dengan tahap data preparation, yaitu penyusunan dataset sintetis menggunakan pendekatan Template-Based Data Generation. Dataset dirancang dalam format instruction–output dengan tujuh kategori intent utama, yakni kpr_simulasi, legal_dokumen, biaya_dan_pajak, spesifikasi_properti, proses_transaksi, konsultasi_keputusan, dan risiko_dan_disclaimer. Setiap data dibuat unik dan bervariasi agar model mampu memahami berbagai skenario percakapan realistis antara calon pembeli dan agen properti. Untuk menjaga kualitas, sebagian data dievaluasi oleh evaluator yang memiliki latar belakang di bidang properti guna menilai relevansi, variasi bahasa, kelengkapan informasi, serta kelayakan jawaban secara praktis.

Setelah dataset dinyatakan layak, data dibagi menjadi set pelatihan dan evaluasi dengan proporsi 90:10 untuk menjaga objektivitas pengujian. Seluruh teks kemudian diproses melalui tokenisizer sebelum dimasukkan ke dalam model.

Tahap fine-tuning dilakukan dengan menyisipkan adaptor LoRA pada lapisan self attention khususnya pada bagian query projection dan value projection sehingga penyesuaian domain dapat dilakukan tanpa memperbarui seluruh parameter model. Konfigurasi pelatihan seperti learning rate, batch size, jumlah epoch, dan mekanisme checkpointing diatur untuk memastikan proses berjalan stabil serta menghindari overfitting maupun catastrophic forgetting. Evaluasi kinerja model dilakukan menggunakan metrik BLEU-4 untuk mengukur kesesuaian tekstual keluaran model terhadap data referensi pada domain properti. Selain evaluasi kuantitatif terhadap kualitas respons, dilakukan pula perbandingan antara chatbot dan agen manusia melalui kuesioner pengguna untuk menilai efektivitas implementasi dari sisi kualitas jawaban dalam konteks layanan konsultasi properti nyata. Secara keseluruhan, penelitian ini memberikan gambaran menyeluruh mulai dari perancangan dataset, proses fine-tuning, evaluasi performa, hingga implementasi sistem chatbot ke dalam antarmuka web yang siap digunakan dalam lingkungan layanan konsultasi properti.

4.2 Dataset Pelatihan

Dataset yang digunakan dalam penelitian ini merupakan kumpulan percakapan berbasis instruction–response yang berfokus pada layanan konsultasi properti. Seluruh data telah melalui proses pembersihan teks, normalisasi bahasa, serta penyesuaian gaya komunikasi agar menyerupai interaksi profesional antara agen properti dan calon pembeli. Respons dalam dataset dirancang informatif, persuasif, dan solutif sehingga tidak hanya memberikan jawaban faktual, tetapi juga mengarahkan pengguna pada langkah lanjutan seperti konsultasi, simulasi

pembiayaan, maupun proses transaksi. Struktur data yang konsisten membuat dataset ini sesuai untuk proses fine-tuning model bahasa menggunakan pendekatan Low-Rank Adaptation, yang bertujuan mengadaptasi model umum menjadi chatbot dengan keahlian khusus di bidang properti.

4.2.1 Template-Based Data Generation

Pada tahap ini, hasil implementasi metode Template-Based Data Generation yang telah dirancang pada bab sebelumnya direalisasikan untuk menghasilkan dataset sintesis chatbot konsultan properti. Proses ini bertujuan untuk memastikan bahwa data yang digunakan dalam fine-tuning benar-benar terstruktur, relevan dengan domain properti, serta memiliki variasi yang memadai untuk melatih model dalam memahami berbagai konteks percakapan. Berbeda dengan pengumpulan data manual yang memerlukan anotasi intensif, pendekatan ini memanfaatkan kemampuan generatif model bahasa untuk menghasilkan pasangan instruction–output secara sistematis berdasarkan template yang telah ditentukan. Hasil dari tahap ini berupa kumpulan data dalam format JSONL yang siap digunakan untuk fine-tuning pada model TinyLLaMA.

Implementasi metode ini dilakukan secara bertahap, dimulai dari penentuan intent, penyusunan format dan prompt, proses generate data dalam jumlah besar, hingga evaluasi kualitas dataset sebelum digunakan dalam pelatihan. Setiap tahapan dirancang untuk menjaga konsistensi struktur data sekaligus memastikan keberagaman skenario percakapan yang dihasilkan.

1. Intent

Pada tahap implementasi intent, tujuh kategori utama yang telah dirancang sebelumnya digunakan sebagai dasar klasifikasi data percakapan, dengan mengacu pada tinjauan topik tanya jawab dalam lingkup agensi properti. Pemilihan ketujuh intent tersebut tidak dilakukan secara acak, melainkan berdasarkan faktor-faktor yang secara teoritis memengaruhi keputusan pembelian properti, seperti aspek finansial, legalitas, kualitas produk, serta risiko transaksi. Hal ini sejalan dengan kajian literatur yang menyatakan bahwa pembelian properti merupakan keputusan kompleks yang melibatkan berbagai pertimbangan ekonomi dan hukum secara simultan (Akhsya et al., 2021; Septiana et al., 2020). Ketujuh intent yang diimplementasikan meliputi `kpr_simulasi`, `legal_dokumen`, `biaya_dan_pajak`, `spesifikasi_properti`, `proses_transaksi`, `konsultasi_keputusan`, serta `risiko_dan_disclaimer`. Masing-masing intent merepresentasikan dimensi penting dalam proses pengambilan keputusan konsumen sebagaimana dijelaskan dalam kajian literatur tersebut.

Intent `kpr_simulasi` berangkat dari kebutuhan konsumen untuk memahami kemampuan finansial dalam membayar cicilan jangka panjang, karena harga dan skema pembiayaan menjadi faktor dominan dalam keputusan pembelian rumah (Akhsya et al., 2021). Intent `legal_dokumen` menitikberatkan pada pentingnya kepastian hukum dalam transaksi, seperti validitas sertifikat dan kelengkapan dokumen, guna meminimalkan risiko sengketa di masa depan (Septiana et al., 2020). Sementara itu, `biaya_dan_pajak` merepresentasikan kebutuhan konsumen untuk mengetahui total beban finansial di luar harga pokok rumah agar tidak terjadi

kesalahan perencanaan anggaran, mengingat biaya tambahan dan pajak merupakan komponen penting dalam transaksi properti (Septiana et al., 2020). Intent spesifikasi_properti berkaitan dengan kualitas bangunan, ukuran, dan fasilitas yang memengaruhi persepsi nilai dan kepuasan konsumen terhadap produk properti (Akhsya et al., 2021). Proses_transaksi mencerminkan tahapan administratif dan legal yang harus dilalui secara transparan sesuai regulasi di Indonesia untuk menjamin keamanan transaksi (Septiana et al., 2020). Adapun konsultasi_keputusan dan risiko_dan_disclaimer merepresentasikan kebutuhan konsumen untuk memperoleh pertimbangan objektif sebelum membeli, termasuk memahami potensi risiko hukum maupun teknis yang mungkin timbul dalam proses pembelian (Akhsya et al., 2021).

Hasil pengelompokan berdasarkan intent menunjukkan bahwa struktur dataset yang dibangun telah selaras dengan kerangka konseptual mengenai faktor-faktor yang memengaruhi keputusan pembelian rumah sebagaimana dibahas dalam penelitian sebelumnya (Akhsya et al., 2021; Septiana et al., 2020). Variasi pertanyaan dalam setiap intent tidak hanya mencerminkan perbedaan profil pengguna (single buyer, pasangan, investor), tetapi juga mencerminkan variasi prioritas pengambilan keputusan, seperti dominasi faktor harga, pertimbangan lokasi dan kualitas produk, maupun kekhawatiran terhadap risiko hukum. Dengan demikian, implementasi intent dalam dataset tidak hanya berfungsi sebagai pengelompokan teknis untuk pelatihan model, tetapi juga sebagai representasi terstruktur dari teori keputusan pembelian properti yang telah dibahas pada tinjauan pustaka.

2. Format dan Prompting

Pada tahap format dan prompting, rancangan struktur dataset yang telah dijelaskan pada Bab III diimplementasikan secara konsisten dalam proses pembuatan data sintetis. Setiap entri data dibangun menggunakan empat atribut utama, yaitu intent, instruction, input, dan output, yang kemudian disimpan dalam format JSON untuk memastikan kompatibilitas dengan proses fine-tuning model. Implementasi format ini bertujuan agar model mampu memahami hubungan antara kategori percakapan, konteks pertanyaan, serta pola jawaban yang diharapkan. Atribut instruction diisi dengan pertanyaan yang disusun menyerupai bahasa alami calon pembeli properti, sedangkan output dirancang dalam bentuk jawaban yang panjang, terstruktur, dan edukatif. Dengan pendekatan ini, model tidak hanya dilatih untuk memberikan respons singkat, tetapi juga mampu menjelaskan langkah perhitungan, asumsi yang digunakan, serta memberikan penegasan atau disclaimer ringan sesuai konteks domain properti.

Dalam tahap prompting, template yang telah dirancang sebelumnya diterapkan untuk menghasilkan variasi percakapan yang realistis dan kontekstual tanpa perlu memaparkan ulang struktur teknisnya. Implementasi aturan variasi dilakukan secara ketat, seperti menghindari pengulangan angka, nominal harga, profil pengguna, maupun pola kalimat yang seragam. Selain itu, setiap respons diwajibkan memiliki kedalaman penjelasan yang memadai agar mencerminkan karakter chatbot sebagai konsultan properti yang informatif dan profesional. Hasil penerapan format dan aturan prompting ini menunjukkan bahwa dataset yang

dihasilkan memiliki keberagaman semantik yang baik, tidak repetitif, serta mampu merepresentasikan berbagai skenario pembelian properti di Indonesia. Dengan demikian, struktur format dan strategi prompting yang diterapkan terbukti mendukung terciptanya dataset yang lebih kaya konteks dan siap digunakan dalam proses fine-tuning model.

3. Generate

Implementasi Pada tahap generate data, template dan format dataset yang telah diimplementasikan pada tahap sebelumnya digunakan sebagai dasar untuk menghasilkan data sintetis secara terstruktur. Proses generasi dilakukan dengan memanfaatkan model GPT untuk menghasilkan variasi percakapan sebanyak ± 3000 entri sesuai dengan tujuh kategori intent yang telah dirancang. Setiap entri dibangun secara konsisten dengan atribut intent, instruction, input, dan output, sehingga tetap selaras dengan kebutuhan fine-tuning model chatbot konsultan properti.

Model menghasilkan pertanyaan yang menyerupai kondisi nyata calon pembeli, seperti variasi penghasilan, harga rumah, skema pembayaran, maupun status pembeli (first-time buyer, pasangan, atau investor). Jawaban yang dihasilkan bersifat analitis, sistematis, dan mencerminkan karakter konsultatif, dengan menyertakan asumsi perhitungan, tahapan penjelasan, serta disclaimer ringan jika diperlukan. Dengan pendekatan ini, dataset tidak hanya kaya secara kuantitas, tetapi juga memiliki kedalaman konteks yang relevan dengan proses pengambilan keputusan pembelian properti.

Penggunaan GPT dalam tahap generasi data memberikan efisiensi signifikan dibandingkan pembuatan manual, karena mampu menghasilkan ribuan variasi percakapan tanpa mengurangi konsistensi struktur. Selain itu, model dapat menjaga keberagaman skenario sehingga dataset tidak repetitif dan tetap representatif terhadap dinamika konsultasi properti di Indonesia. Berikut adalah tiga contoh data hasil proses generate yang merepresentasikan variasi intent berbeda:

Tabel 4. 1 Data Hasil Generate

Atribut	Tipe Data	Contoh 1	Contoh 2
intent	String	"legal dokumen"	"biaya dan pajak"
instruction	String	"Kalau rumah yang saya incar statusnya SHM, apa saja dokumen yang perlu saya cek sebelum transaksi?"	"Selain DP, biaya apa saja yang perlu saya siapkan kalau beli rumah Rp500 juta lewat KPR?"
input	String	""	""
output	String	"Jika statusnya SHM (Sertifikat Hak Milik), pastikan keaslian sertifikat melalui pengecekan di BPN. Periksa kesesuaian nama pemilik, luas tanah, serta pastikan tidak dalam sengketa atau dijaminkan. Selain itu, cek IMB/PBG dan bukti pembayaran PBB terakhir agar tidak ada kewajiban tertunggak."	"Selain DP, Anda perlu menyiapkan BPHTB sekitar 5% dari nilai kena pajak, biaya notaris dan akta 1–2%, biaya provisi bank ±1% dari plafon kredit, serta asuransi jiwa dan kebakaran. Secara umum, total biaya tambahan bisa mencapai 6–10% dari harga rumah, tergantung kebijakan bank dan daerah."

Contoh pada tabel di atas menunjukkan bahwa setiap entri tetap mempertahankan struktur yang konsisten, namun memiliki variasi angka, konteks, dan pendekatan penjelasan. Hal ini membuktikan bahwa proses generate data berhasil menghasilkan dataset yang beragam, realistis, serta siap digunakan dalam tahap fine-tuning model chatbot konsultan properti.

4. Evaluasi Kualitas Dataset

Pada tahap ini, lima responden yang dilibatkan dalam proses evaluasi ini memiliki latar belakang yang berbeda namun relevan dengan konteks properti. Masing-masing responden berasal dari bidang yang mendalami berbagai aspek properti. Tiga di antaranya bekerja di sektor yang langsung berkaitan dengan properti: seorang Agent Properti dari Xavier Marks Premier, seorang Marketing Perumahan dari Primaland, dan seorang Penagih KPR dari Bank BTN, yang memiliki pengalaman dalam hal penyaluran kredit properti. Selain itu, dua responden lainnya berasal dari sektor administratif dan konsultasi: seorang Staff Administrasi di CV. DUTANIAGA KONSTRUKSI, dan seorang Konsultan dari CV. Atha Jaya Konsultan, yang memberikan wawasan dari sisi manajerial dan strategis dalam industri properti.

Dengan latar belakang ini, responden dapat memberikan penilaian yang komprehensif terhadap berbagai kriteria yang diuji, seperti kesesuaian topik, variasi data, keakuratan informasi, dan kelayakan data untuk digunakan dalam pelatihan chatbot properti. Perspektif yang beragam ini memastikan bahwa evaluasi yang dilakukan tidak hanya melihat dari satu sisi, tetapi juga mempertimbangkan berbagai aspek penting dalam transaksi properti dan konsultasi yang dapat terjadi di dunia nyata.

Tabel 4. 2 Penilaian Kualitas dan Relevansi Dataset Chatbot Properti

Kriteria Evaluasi	Responden 1	Responden 2	Responden 3	Responden 4	Responden 5
Topik yang Diberikan	Simulasi KPR, Legal Dokumen, Biaya dan Pajak, Proses Transaksi,	Simulasi KPR, Legal Dokumen, Biaya dan Pajak, Proses Transaksi,	Simulasi KPR, Legal Dokumen, Biaya dan Pajak, Proses	Simulasi KPR, Legal Dokumen, Biaya dan Pajak, Proses	Simulasi KPR, Legal Dokumen, Biaya dan Pajak, Proses

	Konsultasi Keputusan, Risiko dan Disclaimer, Spesifikasi Properti	Konsultasi Keputusan, Risiko dan Disclaimer, Spesifikasi Properti	Transaksi, Konsultasi Keputusan, Risiko dan Disclaimer , Spesifikasi Properti	Transaksi, Konsultasi Keputusan, Risiko dan Disclaimer, Spesifikasi Properti	Transaksi, Konsultasi Keputusan , Risiko dan Disclaimer , Spesifikasi Properti
Variasi Data	Cukup Bervariasi	Sangat Bervariasi	Sangat Bervariasi	Cukup Bervariasi	Sangat Bervariasi
Realitas dan Kesesuaian Konteks	Sangat realistis dan sesuai	Sangat realistis dan sesuai	Sangat realistis dan sesuai	Sangat realistis dan sesuai	Sangat realistis dan sesuai
Akurasi dan Bebas dari Kesalahan	Sangat Akurat	Sangat Akurat	Sangat Akurat	Sangat Akurat	Sangat Akurat
Keluwes dan Kejelasan	Baik	Baik	Sangat Baik	Baik	Sangat Baik
Kepahama n oleh Pengguna	Sangat Mudah Dipahami	Sangat Mudah Dipahami	Cukup Mudah Dipahami	Sangat Mudah Dipahami	Sangat Mudah Dipahami
Kesesuaian Tujuan	Sangat mudah	Sangat mudah	Sangat mudah	Sangat mudah	Sangat mudah
Kesalahan atau Ketidaks esuaian	Tidak ada	Tidak menemukan, semua lengkap	Saya rasa sudah bisa dipahami	-	Tidak ada
Saran untuk Peningkata n	Sudah Bagus	Perlu dikembangkan lagi, agar wawasan tentang ilmu dan informasi apapun selalu update	Mungkin asumsi di membeli perumahan dengan minim transum bisa ditambahkan lagi	-	Sudah sangat bagus

Berdasarkan hasil evaluasi, seluruh responden sepakat bahwa topik-topik dalam dataset seperti Simulasi KPR, Legal Dokumen, Biaya dan Pajak, serta topik lainnya sudah cukup relevan dan sesuai dengan konteks yang dimaksud, yaitu konsultasi properti. Semua responden menilai bahwa dataset ini sangat realistis dan sesuai dengan kebutuhan domain properti.

Dari sisi variasi data, sebagian besar responden (terutama Responden 2, 3, dan 5) memberikan penilaian positif bahwa variasi data yang ada sudah sangat bervariasi. Namun, Responden 1 dan 4 sedikit kurang puas, karena menganggap variasi data masih cukup bervariasi, yang menunjukkan masih ada ruang untuk meningkatkan keragaman dalam skenario percakapan yang dihadirkan.

Terkait akurasi dan bebas dari kesalahan, seluruh responden sepakat bahwa data yang disediakan sangat akurat dan bebas dari kesalahan yang jelas, yang memastikan kualitas informasi yang diberikan sangat tinggi. Aspek keluwesan dan kejelasan informasi juga mendapat nilai baik, meskipun Responden 3 memberikan penilaian terbaik dengan menyebutkan bahwa data memiliki keluwesan dan kejelasan yang sangat baik.

Mengenai kepastian oleh pengguna, sebagian besar responden menganggap data yang ada sangat mudah dipahami, dengan satu responden yang menyebutkan bahwa pemahaman data cukup mudah dipahami, namun tidak mempengaruhi penilaian secara keseluruhan.

Pada bagian kesesuaian tujuan, seluruh responden sepakat bahwa dataset ini sangat sesuai untuk tujuan konsultasi properti dan perencanaan pembelian rumah, karena data yang disediakan dapat digunakan untuk memberikan jawaban yang relevan sesuai dengan tujuan yang dimaksud.

Terakhir, terkait dengan kesalahan atau ketidaksesuaian, sebagian besar responden tidak menemukan kesalahan besar dalam dataset. Beberapa responden memberikan saran untuk meningkatkan variasi dan pengembangan skenario lebih lanjut, dengan Responden 2 memberi saran agar wawasan mengenai informasi

selalu diperbarui, serta menambahkan asumsi terkait pembelian rumah dengan transaksi minim.

Secara keseluruhan, evaluasi ini menunjukkan bahwa dataset yang digunakan untuk fine-tuning model sudah memenuhi standar kualitas yang cukup dalam hal relevansi, keakuratan, dan pemahaman oleh pengguna, meskipun ada beberapa area yang dapat diperbaiki dengan penambahan variasi skenario dan pembaruan informasi lebih lanjut.

4.2.2 Ringkasan Akhir Dataset

Dataset akhir yang digunakan berjumlah 3024 pasangan data instruction–output. Setiap data merepresentasikan satu skenario tanya jawab antara pengguna dan agen properti virtual. Distribusi topik dalam dataset mencakup berbagai aspek penting dalam proses pembelian properti, mulai dari simulasi pembiayaan hingga aspek legal dan risiko transaksi. Keberagaman topik ini mendukung pembelajaran model agar mampu memberikan respons yang komprehensif dalam berbagai situasi layanan. Berikut ringkasan distribusi topik utama dalam dataset:

Tabel 4. 3 Tabel Jumlah Intent pada Dataset

Kategori Intent	Jumlah Data	Deskripsi Singkat
KPR Simulasi	575	Perhitungan cicilan, DP, tenor, dan kelayakan kredit
Legal & Dokumen	576	Informasi sertifikat, IMB, AJB, dan legalitas properti
Biaya & Pajak	504	Perkiraan biaya tambahan, pajak, notaris, dan administrasi
Konsultasi Keputusan	402	Saran memilih properti sesuai kondisi finansial pengguna
Spesifikasi Properti	331	Detail fisik rumah seperti luas, kamar, dan fasilitas
Risiko & Disclaimer	328	Penjelasan risiko transaksi dan batasan

		informasi
Proses Transaksi	326	Tahapan pembelian, booking, hingga akad kredit

Selain distribusi topik yang luas, dataset juga menggunakan format instruction–response. Contoh struktur data adalah sebagai berikut:

Tabel 4. 4 Format Dataset

Instruction	Output
“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?”	“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.”

Format ini membantu model mempelajari hubungan antara kebutuhan pengguna dan respons yang relevan serta berorientasi layanan. Dengan jumlah data yang cukup, cakupan topik yang beragam, dan gaya bahasa yang konsisten, dataset ini dinilai memadai untuk proses fine-tuning LoRA guna membentuk chatbot properti yang responsif dan kontekstual.

4.3 Pelatihan Model

Proses pelatihan pada penelitian ini dilakukan melalui beberapa tahapan, dimulai dari konfigurasi pelatihan model, pembahasan hasil fine-tuning pada tiga skenario, hingga evaluasi kualitas respons model untuk menentukan model terbaik. Setiap skenario menggunakan dataset dan arsitektur yang sama dengan variasi pada LoRA Rank (r) dan LoRA Alpha (α) sebagai variabel utama, sementara

hyperparameter lainnya dikontrol secara konsisten agar perbedaan hasil pelatihan dapat diatribusikan pada kedua parameter tersebut secara terisolasi.

Seluruh proses pelatihan dilakukan menggunakan GPU T4 dengan pembagian data sebesar 90% untuk pelatihan dan 10% untuk validasi. LoRA Dropout ditetapkan pada nilai 0 di seluruh skenario mempertimbangkan ukuran dataset yang terbatas sekitar 3024 sampel, sehingga penerapan dropout berpotensi menyebabkan underfitting selama pelatihan. Learning rate sebesar 2×10^{-4} dan jumlah epoch sebesar 12 dikontrol secara konsisten di seluruh skenario agar perbandingan hasil pelatihan dapat diatribusikan pada variasi rank dan alpha secara terisolasi.

4.3.1 Konfigurasi Pelatihan Model

Model dasar yang digunakan adalah TinyLLaMA-1.1B-Chat-v1.0, yang kemudian diadaptasi menggunakan teknik Low-Rank Adaptation (LoRA) pada layer query projection (q_proj) dan value projection (v_proj). Pada setiap eksperimen, parameter LoRA seperti rank (r) dan alpha diatur untuk mengontrol kapasitas adaptasi model terhadap domain properti. Proses pelatihan dilakukan dengan menggunakan optimizer AdamW.

Setiap skenario pelatihan menggunakan per-device batch size sebesar 4 dengan gradient accumulation steps sebesar 2, sehingga menghasilkan effective batch size sebesar 8. Jumlah epoch untuk semua skenario sebanyak 12 epoch. Pengaturan ini dikontrol secara konsisten agar perbedaan hasil pelatihan tidak dipengaruhi oleh variasi kapasitas input, melainkan murni berasal dari perbedaan nilai LoRA Rank (r) dan Alpha (α) yang menjadi variabel utama eksperimen.

Sebagai bagian dari proses eksperimen, dilakukan perbandingan tiga skenario konfigurasi dengan nilai rank $r = 8$, $r = 16$, dan $r = 32$ guna mengevaluasi pengaruh variasi kapasitas adaptasi terhadap performa model.

1. Skenario 1

Skenario 1 menggunakan LoRA rank sebanyak 8 dan alpha sebanyak 16, menjadikannya konfigurasi dengan kapasitas adaptasi paling rendah di antara ketiga skenario. Dengan nilai rank yang kecil, jumlah parameter adaptif yang dilatih lebih sedikit sehingga kemampuan model dalam menangkap pola spesifik domain properti pada skenario ini relatif lebih terbatas dibandingkan skenario lainnya.

Tabel 4. 5 Konfigurasi Pelatihan Skenario 1

Parameter	Konfigurasi LoRA pada Skenario 1
Model Dasar	TinyLLaMA-1.1B-Chat-v1.0
LoRA Rank (r)	8
LoRA Alpha	16
LoRA Dropout	0
Target Layer	q_{proj}, v_{proj}
Learning Rate	2×10^{-4}
Epoch Maksimum	12
Batch Size Efektif	8
Batch Size (Train)	4
Gradient Accumulation	2
Max Sequence Length	1024
Scheduler	Cosine
Optimizer	AdamW
Max Grad Norm	1.0
Warmup Ratio	0.05

2. Skenario 2

Skenario 2 menggunakan LoRA rank sebanyak 16 dan alpha sebanyak 32, merepresentasikan peningkatan kapasitas adaptasi sebesar dua kali lipat dibandingkan Skenario 1. Hu et al. (2021) menunjukkan bahwa peningkatan rank menghasilkan ruang representasi yang lebih kaya, sehingga pada skenario ini model

beroperasi dengan kapasitas yang lebih besar dalam mempelajari pola instruksi dan terminologi domain properti, dengan seluruh hyperparameter lainnya dikontrol identik agar perbedaan hasil dapat diatribusikan murni pada variasi rank dan alpha.

Tabel 4. 6 Konfigurasi Pelatihan Skenario 2

Parameter	Konfigurasi LoRA pada Skenario 2
Model Dasar	TinyLLaMA-1.1B-Chat-v1.0
LoRA Rank (r)	16
LoRA Alpha	32
LoRA Dropout	0
Target Layer	q_proj, v_proj
Learning Rate	2×10^{-4}
Epoch Maksimum	12
Batch Size Efektif	8
Batch Size (Train)	4
Gradient Accumulation	2
Max Sequence Length	1024
Scheduler	Cosine
Optimizer	AdamW
Max Grad Norm	1.0
Warmup Ratio	0.05

3. Skenario 3

Skenario 3 menggunakan LoRA rank sebanyak 32 dan alpha sebanyak 64 sebagai konfigurasi dengan kapasitas adaptasi tertinggi dalam skenario. Nilai rank yang lebih besar menghasilkan ruang representasi yang lebih luas dibandingkan dua skenario sebelumnya, sehingga skenario ini digunakan untuk menguji apakah penambahan kapasitas parameter adaptif secara signifikan berpengaruh terhadap kualitas respons model pada domain properti.

Tabel 4. 7 Konfigurasi Pelatihan Skenario 3

Parameter	Konfigurasi LoRA pada Skenario 3
Model Dasar	TinyLLaMA-1.1B-Chat-v1.0
LoRA Rank (r)	32
LoRA Alpha	64
LoRA Dropout	0
Target Layer	q_proj, v_proj
Learning Rate	2×10^{-4}
Epoch Maksimum	12
Batch Size Efektif	8
Batch Size (Train)	4

Gradient Accumulation	2
Max Sequence Length	1024
Scheduler	Cosine
Optimizer	AdamW
Max Grad Norm	1.0
Warmup Ratio	0.05

4.3.2 Hasil Proses Pelatihan Model

Proses pelatihan dilakukan untuk mengevaluasi pengaruh variasi LoRA Rank dan Alpha terhadap stabilitas pembelajaran dan performa model pada domain properti. Analisis hasil pelatihan difokuskan pada grafik Training Loss vs Step untuk mengamati dinamika penurunan loss pada setiap langkah pembaruan parameter serta mengidentifikasi pola konvergensi dan potensi fluktuasi loss antar skenario.

Perbandingan antar skenario dilakukan dengan mengaitkan pola grafik terhadap perbedaan nilai rank $r = 8$, $r = 16$, dan $r = 32$, sementara hyperparameter lainnya dikontrol identik di seluruh skenario. Dengan pendekatan ini, evaluasi tidak hanya berfokus pada nilai loss akhir, tetapi juga pada dinamika proses pembelajaran, sehingga dapat diidentifikasi nilai rank dan alpha yang menghasilkan konvergensi paling stabil dalam mengadaptasi model ke domain properti.

1. Skenario 1

Hasil pelatihan Skenario 1 dengan konfigurasi LoRA Rank $r = 8$ dan Alpha $\alpha = 16$ menunjukkan pola konvergensi yang stabil sepanjang proses pelatihan sebagaimana terlihat pada Gambar 4.1. Pada fase awal pelatihan, training loss dimulai dari nilai sekitar 2.2 dan mengalami penurunan yang tajam hingga step ke-500, yang mengindikasikan bahwa model secara cepat menyesuaikan parameter

adaptif LoRA terhadap distribusi data domain properti pada tahap awal optimasi. Penurunan yang tajam ini konsisten dengan karakteristik cosine scheduler yang menerapkan warmup ratio 0.05 pada fase awal, sehingga laju pembelajaran meningkat secara bertahap sebelum mencapai nilai puncaknya.



Gambar 4. 1 Grafik Training Loss vs Step Skenario 1

Memasuki rentang step 500 hingga 1500, laju penurunan loss melambat dan disertai fluktuasi yang terlihat pada beberapa titik pada Gambar 4.1. Fluktuasi ini merupakan pola yang wajar pada proses fine-tuning dengan effective batch size sebesar 8, di mana estimasi gradien per langkah pembaruan belum sepenuhnya merepresentasikan keseluruhan distribusi data. Pada rentang step 1500 hingga 3000, penurunan loss berlanjut secara lebih stabil dengan amplitudo fluktuasi yang semakin mengecil, menandakan bahwa model mulai memasuki fase konvergensi. Pada akhir pelatihan di sekitar step 4300, training loss mencapai nilai sekitar 1.05 sebagaimana ditunjukkan pada Gambar 4.1, dengan tren yang masih menunjukkan penurunan. Nilai rank $r = 8$ yang relatif kecil menghasilkan jumlah parameter

adaptif yang terbatas, sehingga model pada skenario ini beroperasi dengan kapasitas representasi yang lebih sempit dibandingkan skenario berikutnya.

2. Skenario 2

Hasil pelatihan Skenario 2 dengan konfigurasi LoRA Rank $r = 16$ dan Alpha $\alpha = 32$ menunjukkan pola konvergensi yang lebih dalam dibandingkan Skenario 1 sebagaimana terlihat pada Gambar 4.2. Pada fase awal pelatihan, training loss dimulai dari nilai sekitar 2.18 dan mengalami penurunan tajam hingga step ke-500, dengan laju penurunan yang sebanding dengan Skenario 1 pada fase yang sama. Pola ini konsisten dengan penerapan cosine scheduler dan warmup ratio 0.05 yang identik antar skenario, sehingga perbedaan perilaku pelatihan pada fase berikutnya dapat diatribusikan pada perbedaan nilai rank dan alpha.



Gambar 4. 2 Grafik Training Loss vs Step Skenario 2

Memasuki rentang step 500 hingga 1500, penurunan loss pada Skenario 2 terlihat lebih konsisten dengan fluktuasi yang relatif lebih terkendali dibandingkan Skenario 1 pada rentang step yang sama sebagaimana ditunjukkan pada Gambar 4.2. Pola ini mengindikasikan bahwa peningkatan nilai rank dari $r = 8$ ke $r = 16$ berdampak pada jumlah parameter adaptif yang lebih besar, sehingga proses penyesuaian bobot terhadap distribusi data domain properti berlangsung lebih stabil. Pada rentang step 1500 hingga 3000, penurunan loss berlanjut secara stabil hingga mendekati nilai 0.9. Pada akhir pelatihan di sekitar step 4300, training loss mencapai nilai sekitar 0.88 sebagaimana ditunjukkan pada Gambar 4.2, lebih rendah dibandingkan nilai akhir Skenario 1 yang berada di sekitar 1.05, dengan tren loss yang semakin melandai menjelang akhir pelatihan.

3. Skenario 3

Hasil pelatihan Skenario 3 dengan konfigurasi LoRA Rank $r = 32$ dan Alpha $\alpha = 64$ menunjukkan pola konvergensi yang paling dalam di antara ketiga skenario sebagaimana terlihat pada Gambar 4.3. Pada fase awal pelatihan, training loss dimulai dari nilai sekitar 2.15 dan mengalami penurunan tajam hingga step ke 500, dengan laju penurunan yang sebanding dengan Skenario 1 dan Skenario 2 pada fase yang sama. Kesamaan pola pada fase awal ini konsisten dengan penggunaan cosine scheduler dan warmup ratio 0.05 yang identik di seluruh skenario, sehingga perbedaan perilaku pelatihan pada fase berikutnya dapat diatribusikan pada perbedaan nilai rank dan alpha.



Gambar 4. 3 Grafik Training Loss vs Step Skenario 3

Memasuki rentang step 500 hingga 1500, penurunan loss pada Skenario 3 terlihat lebih landai dan konsisten dibandingkan dua skenario sebelumnya sebagaimana ditunjukkan pada Gambar 4.3, tanpa fluktuasi yang tajam pada rentang step tersebut. Pada rentang step 1500 hingga 3000, penurunan loss berlanjut secara stabil hingga mendekati nilai 0.75, lebih rendah dibandingkan Skenario 1 yang berada di sekitar 1.05 dan Skenario 2 di sekitar 0.9 pada rentang step yang sama. Pola ini mengindikasikan bahwa nilai rank $r = 32$ dengan dimensi matriks dekomposisi AB yang lebih besar menghasilkan jumlah parameter adaptif yang lebih banyak dibandingkan dua skenario sebelumnya, sehingga proses adaptasi terhadap pola spesifik domain properti berlangsung lebih dalam sepanjang proses pelatihan. Pada akhir pelatihan di sekitar step 4300, training loss mencapai nilai sekitar 0.67 sebagaimana ditunjukkan pada Gambar 4.3, dengan tren loss yang semakin melandai dan fluktuasi yang mengecil menjelang akhir pelatihan.

4.3.3 Evaluasi Kualitas Respons Model

Evaluasi kualitas respons model dilakukan untuk mengukur dan membandingkan kinerja ketiga skenario model hasil fine-tuning dengan variasi LoRA Rank dan Alpha yang berbeda pada masing-masing skenario. Penilaian kinerja dilakukan menggunakan metrik BLEU-4 untuk mengukur kesesuaian keluaran model terhadap data referensi. Melalui pendekatan evaluasi ini, perbandingan antar skenario dapat dilakukan secara konsisten berdasarkan tingkat kesesuaian tekstual jawaban model terhadap referensi dalam domain properti.

1. Skenario 1

Evaluasi kualitas respons Skenario 1 dilakukan terhadap model hasil fine-tuning dengan konfigurasi LoRA Rank $r = 8$ dan Alpha $\alpha = 16$ menggunakan metrik BLEU-4.

Tabel 4. 8 Evaluasi Metriks Skenario 1

Metriks	Skor
BLEU-4	12.76

Skenario 1 menghasilkan skor BLEU-4 sebesar 12.76 sebagaimana ditunjukkan pada Tabel 4.8, yang merupakan nilai terendah di antara ketiga skenario. Skor ini mengindikasikan bahwa dengan nilai rank yang paling kecil, jumlah parameter adaptif yang tersedia selama proses fine-tuning lebih terbatas, sehingga kesesuaian keluaran model terhadap data referensi pada domain properti berada pada tingkat yang lebih rendah dibandingkan skenario dengan rank yang lebih besar.

Hasil evaluasi pada Tabel 4.8 menunjukkan bahwa konfigurasi LoRA Rank $r = 8$ menghasilkan kesesuaian tekstual yang lebih rendah terhadap referensi dibandingkan dua skenario lainnya, yang sejalan dengan pola training loss akhir Skenario 1 yang juga berada pada nilai 1.05.

2. Skenario 2

Evaluasi kualitas respons Skenario 2 dilakukan terhadap model hasil fine-tuning dengan konfigurasi LoRA Rank $r = 16$ dan Alpha $\alpha = 32$ menggunakan metrik BLEU-4.

Tabel 4. 9 Evaluasi Metriks Skenario 2

Metriks	Skor
BLEU-4	14.49

Skenario 2 menghasilkan skor BLEU-4 sebesar 14.49 sebagaimana ditunjukkan pada Tabel 4.9, lebih tinggi dibandingkan Skenario 1 dengan selisih sekitar 1.73 poin. Peningkatan ini sejalan dengan peningkatan nilai rank dari $r = 8$ ke $r = 16$ yang menghasilkan jumlah parameter adaptif lebih besar, sehingga model pada skenario ini menghasilkan keluaran yang lebih sesuai dengan data referensi.

Hasil evaluasi pada Tabel 4.9 menunjukkan bahwa peningkatan LoRA Rank dari $r = 8$ ke $r = 16$ berdampak pada peningkatan kesesuaian tekstual, dengan pola yang konsisten terhadap training loss akhir Skenario 2 yang lebih rendah dibandingkan Skenario 1, yaitu berada di sekitar nilai 0.88.

3. Skenario 3

Evaluasi kualitas respons Skenario 3 dilakukan terhadap model hasil fine-tuning dengan konfigurasi LoRA Rank $r = 32$ dan Alpha $\alpha = 64$ menggunakan metrik BLEU-4.

Tabel 4. 10 Evaluasi Metriks Skenario 3

Metriks	Skor
BLEU-4	14.54

Skenario 3 menghasilkan skor BLEU-4 sebesar 14.54 sebagaimana ditunjukkan pada Tabel 4.10, merupakan nilai tertinggi di antara ketiga skenario namun hanya selisih 0.05 poin dibandingkan Skenario 2. Perbedaan yang sangat kecil ini mengindikasikan bahwa peningkatan rank dari $r = 16$ ke $r = 32$ tidak menghasilkan perbedaan kesesuaian tekstual yang substansial, meskipun training loss akhir Skenario 3 berada lebih rendah di sekitar 0.67 dibandingkan Skenario 2.

Hasil evaluasi pada Tabel 4.10 menunjukkan bahwa penambahan kapasitas rank dari $r = 16$ ke $r = 32$ menghasilkan peningkatan skor BLEU-4 yang tidak signifikan, sehingga kenaikan nilai rank pada rentang tersebut tidak secara proporsional berdampak pada peningkatan kesesuaian keluaran model terhadap data referensi dalam domain properti.

4.3.4 Analisis Model Terbaik

Dalam penelitian ini, dilakukan serangkaian evaluasi terhadap tiga skenario model yang telah dilatih menggunakan pendekatan Low-Rank Adaptation pada model dasar TinyLLaMA-1.1B-Chat-v1.0 dengan variasi LoRA Rank dan Alpha sebagai parameter pembeda. Hasil evaluasi menunjukkan bahwa setiap skenario

menghasilkan tingkat kesesuaian tekstual yang berbeda terhadap data referensi domain properti, sebagaimana diukur menggunakan metrik BLEU-4.

Skenario 1 dengan konfigurasi LoRA $r = 8$ dan $\alpha = 16$ menghasilkan skor BLEU-4 sebesar 12.76, merupakan nilai terendah di antara ketiga skenario. Hal ini sejalan dengan training loss akhir Skenario 1 yang masih cukup tinggi di angka 1.05, menunjukkan bahwa rank yang kecil membuat model kurang mampu menyesuaikan keluarannya dengan data referensi.

Skenario 2 dengan konfigurasi LoRA $r = 16$ dan $\alpha = 32$ menghasilkan skor BLEU-4 sebesar 14.49, lebih tinggi 1.73 poin dibandingkan Skenario 1. Peningkatan ini sejalan dengan training loss akhir Skenario 2 yang lebih rendah pada rentang angka 0.8, menunjukkan bahwa penambahan rank dari $r = 8$ ke $r = 16$ memberikan dampak yang cukup nyata terhadap kualitas keluaran model.

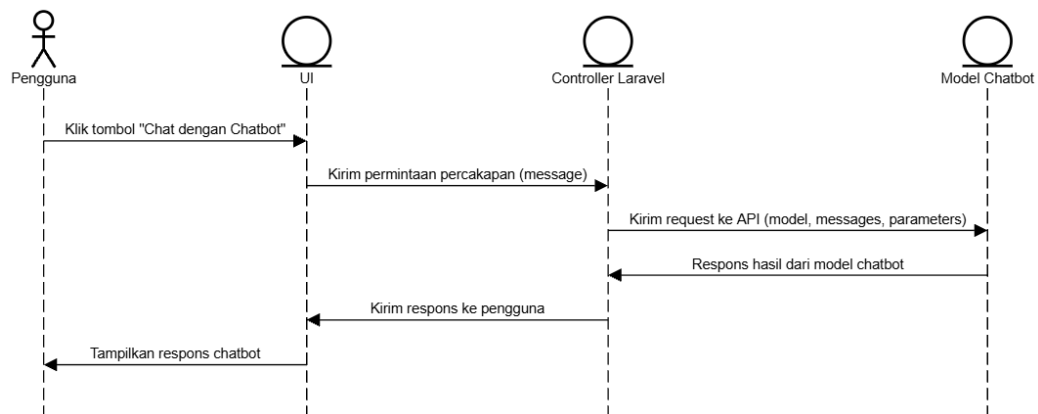
Skenario 3 dengan konfigurasi LoRA $r = 32$ dan $\alpha = 64$ menghasilkan skor BLEU-4 tertinggi yaitu 14.54, namun hanya unggul 0.05 poin dibandingkan Skenario 2. Meskipun training loss akhir Skenario 3 merupakan yang terendah pada sekitar rentang angka 0.67, selisih skor BLEU-4 yang sangat kecil menunjukkan bahwa penambahan rank dari $r = 16$ ke $r = 32$ tidak memberikan peningkatan yang berarti pada kualitas keluaran model.

Berdasarkan hasil evaluasi keseluruhan, Skenario 3 mencatat skor BLEU-4 tertinggi diantara skenario lainnya dan dipilih sebagai model terbaik untuk diimplementasikan pada sistem chatbot konsultan properti. Meskipun demikian, selisih yang sangat kecil antara Skenario 2 dan Skenario 3 menunjukkan bahwa

peningkatan rank di atas $r = 16$ tidak secara proporsional meningkatkan kualitas keluaran model pada domain properti dengan ukuran dataset sekitar 3024 sampel.

4.4 Implementasi Model Pada Sistem

Pada tahap implementasi ini, model chatbot konsultan properti yang telah dilatih dengan pendekatan Low-Rank Adaptation menggunakan model TinyLLaMA diintegrasikan dengan sistem berbasis web menggunakan framework Laravel. Integrasi ini bertujuan untuk memberikan layanan chatbot yang dapat diakses pengguna melalui antarmuka web secara langsung. Sistem ini dijalankan pada perangkat dengan spesifikasi Intel Core i5 Gen 13420H, GPU RTX 4050 dengan VRAM 6GB, dan RAM 8GB, yang digunakan untuk menjalankan model chatbot serta sistem web berbasis Laravel.



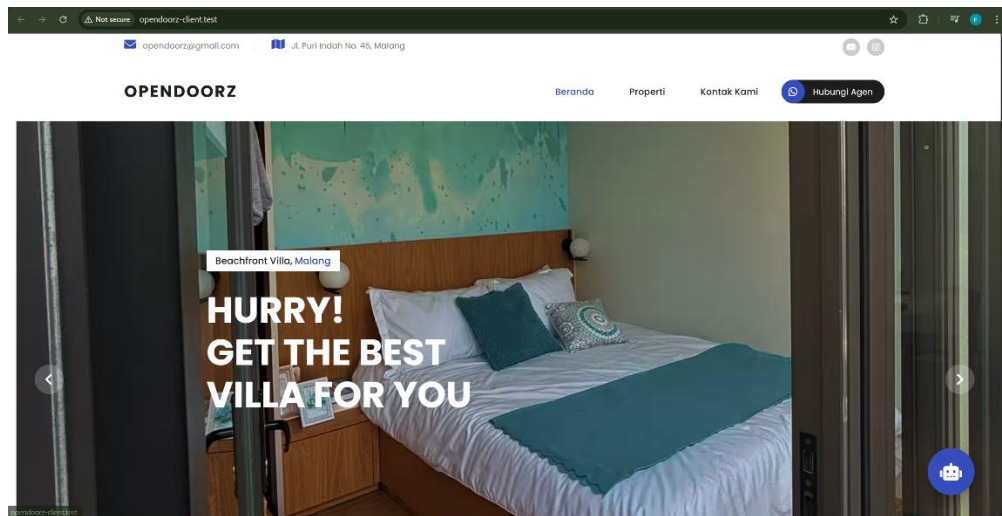
Gambar 4. 4 Sequence Diagram Alur Komunikasi antara Pengguna dengan Chatbot

Sequence diagram pada Gambar 4.4 menggambarkan alur interaksi pengguna dengan sistem chatbot yang telah diimplementasikan. Berbeda dengan desain teknis pada BAB III yang berfokus pada mekanisme streaming antar komponen, diagram ini menampilkan perspektif pengguna secara menyeluruh

mulai dari inisiasi percakapan hingga penerimaan respons chatbot. Detail mekanisme streaming SSE telah dijelaskan secara lengkap pada BAB III sehingga tidak diulang pada bagian ini.

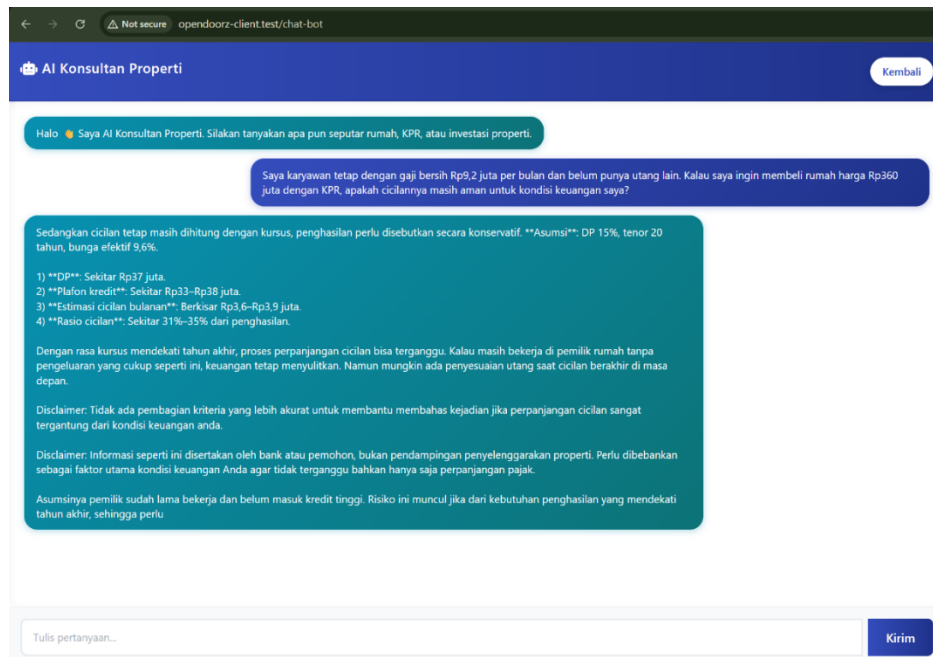
Alur interaksi dimulai ketika pengguna mengklik tombol "Chat dengan Chatbot" yang tersedia pada antarmuka pengguna. Tindakan ini memicu pengiriman permintaan percakapan dari UI ke Controller Laravel. Selanjutnya, controller meneruskan permintaan tersebut ke API yang menghubungkan sistem dengan model chatbot. Model chatbot kemudian memproses permintaan dan menghasilkan respons sesuai dengan input pengguna. Setelah respons diterima, controller mengirimkannya kembali ke UI untuk ditampilkan kepada pengguna, sehingga menyelesaikan satu siklus percakapan.

Untuk memfasilitasi komunikasi antara sistem frontend dan model backend, digunakan API yang dapat dipanggil melalui controller Laravel. Kode controller berikut mengirimkan permintaan ke endpoint, yang mengirimkan pesan pengguna dan menunggu respons dari model chatbot. Pesan yang dikirim berisi instruksi sistem yang mengonfigurasi chatbot sebagai konsultan properti, serta pesan dari pengguna yang berisi pertanyaan mengenai properti, pembiayaan rumah, atau investasi properti.



Gambar 4. 5 Landing Page Website

Sistem frontend menampilkan tombol interaksi chatbot di pojok kanan bawah layar sebagai titik masuk pengguna untuk memulai percakapan. Setelah tombol diklik, jendela percakapan muncul, dan pengguna dapat mulai bertanya tentang berbagai topik terkait properti. Di sisi server, model TinyLLaMA yang telah dilatih dengan data spesifik domain properti memberikan respons yang sesuai dengan konteks pertanyaan yang diberikan. Respons ini dihasilkan berdasarkan model yang telah diproses dengan dataset yang berfokus pada topik-topik seperti simulasi KPR, dokumen legal, biaya, dan pajak properti, serta konsultasi keputusan pembelian properti.



Gambar 4. 6 Tampilan UI Chat antara User dengan Chatbot

Selama percakapan berlangsung, chatbot memberikan jawaban dari pertanyaan yang diajukan oleh pengguna. Misalnya, ketika pengguna bertanya tentang kemungkinan pembelian rumah dengan pembiayaan KPR, chatbot memberikan perhitungan yang jelas mengenai cicilan, bunga, dan estimasi biaya lainnya. Pada tahap implementasi ini, chatbot memberikan penjelasan berdasarkan informasi yang tersedia dalam dataset domain properti, mencakup topik simulasi KPR, dokumen legal, biaya, pajak properti, dan konsultasi keputusan pembelian.

4.5 Uji Implementasi Model

Proses pengujian dilakukan untuk membandingkan performa chatbot berbasis TinyLLaMA yang telah di fine-tune dengan LoRA dan agen manusia dalam konteks layanan properti. Responden berinteraksi dengan salah satu dari tujuh topik utama, kemudian mengisi kuesioner untuk menilai kualitas jawaban,

kepuasan, preferensi sistem. Struktur kuesioner lengkap dapat dilihat pada Tabel 3.8 pada Bab 3. Pendekatan ini memungkinkan pengumpulan data secara terstruktur, sekaligus menilai pengalaman pengguna secara relevan dalam konteks layanan properti.

4.5.1 Pembahasan Hasil Uji Implementasi Model

Subbab ini membahas hasil uji implementasi model chatbot berbasis TinyLLaMA dengan fine-tuning LoRA dari sisi persepsi pengguna, yang diperoleh melalui penyebaran kuesioner skala Likert 1–5 kepada 45 responden. Pengujian ini mencakup empat variabel pengukuran, yaitu kelengkapan jawaban, kemudahan pemahaman, kepuasan terhadap chatbot, dan kepuasan terhadap agen manusia sebagai pembanding. Hasil pengolahan data berupa skor total, rata-rata yang nantinya digunakan sebagai dasar untuk melihat kecenderungan persepsi responden terhadap kualitas jawaban dan tingkat kepuasan yang dihasilkan oleh model pada tahap pengujian ini. Data hasil pengujian model oleh pengguna yang dilakukan oleh 45 responden disajikan pada Tabel 4.11 berikut.

Tabel 4. 11 Data Hasil Kuesioner Skala Likert per Responden

No	Kelengkapan Jawaban	Kemudahan Pemahaman	Kepuasan Chatbot	Kepuasan Agen Manusia
1	3	5	3	4
2	4	3	2	5
3	1	1	1	1
4	1	2	2	1
5	1	1	1	1
6	1	1	1	1
7	1	4	3	2
8	1	1	1	1
9	4	3	3	2
10	3	3	4	3
11	4	3	3	3

12	1	1	1	1
13	4	3	3	4
14	4	3	4	3
15	2	2	1	3
16	4	3	3	3
17	3	3	3	3
18	1	1	1	1
19	3	2	4	3
20	3	3	3	3
21	4	4	4	4
22	3	3	3	3
23	3	3	3	3
24	2	2	2	2
25	2	3	3	3
26	5	3	4	2
27	2	2	2	2
28	3	3	3	3
29	4	4	4	4
30	1	1	1	1
31	3	4	4	4
32	1	1	2	1
33	3	4	4	3
34	3	3	3	3
35	3	3	3	3
36	3	4	3	5
37	5	5	5	5
38	2	2	2	2
39	4	5	5	4
40	3	4	4	4
41	4	4	3	4
42	3	3	3	3
43	4	4	4	3
44	3	3	2	3
45	4	4	3	4

Berdasarkan dari Tabel 4.11 hasil data dari skala likert selanjutnya dirangkum ke dalam bentuk statistik deskriptif sebagaimana yang disajikan pada Tabel 4.12.

Tabel 4. 12 Statistik Deskriptif Empat Variabel Pengukuran

Variabel	Skor Total	Rata-rata	Interpretasi Rata-rata	Min	Max
Kelengkapan Jawaban	126	2,80	Cukup Lengkap	1	5
Kemudahan Pemahaman	129	2,87	Cukup Mudah Dipahami	1	5
Kepuasan Chatbot	126	2,80	Cukup Puas	1	5
Kepuasan Agen Manusia	126	2,80	Cukup Puas	1	5

Penentuan kategori kualitas pada Tabel 4.12 didasarkan pada pedoman interval skor sebagaimana disajikan pada Tabel 4.13.

Tabel 4. 13 Pedoman Interval Skor Skala Likert

Interval Skor	Keterangan
1.00 – 1.80	Sangat Baik / Sangat Puas
1.81 – 2.60	Baik / Puas
2.61 – 3.40	Cukup
3.41 – 4.20	Buruk / Tidak Puas
4.21 – 5.00	Sangat Buruk / Sangat Tidak Puas

Berdasarkan Tabel 4.12, hasil pengujian implementasi model terhadap persepsi pengguna menunjukkan bahwa keempat variabel yang diukur, yaitu kelengkapan jawaban, kemudahan pemahaman, kepuasan terhadap chatbot, dan kepuasan terhadap agen manusia, secara konsisten berada pada rentang nilai rata-rata 2.80 hingga 2.87. Berdasarkan pedoman interval skor pada Tabel 4.13, keseluruhan nilai tersebut termasuk dalam kategori Cukup (2.61–3.40), yang menunjukkan bahwa performa model belum mencapai kategori baik, namun juga tidak berada pada kategori buruk.

Pada variabel kelengkapan jawaban, hasil rekapitulasi pada Tabel 4.12 menunjukkan skor total sebesar 126 dengan rata-rata 2.80, yang berada pada

interpretasi "Cukup Lengkap". Nilai rata-rata tersebut berada pada rentang skor minimum 1 dan maksimum 5, menunjukkan bahwa secara keseluruhan jawaban yang dihasilkan model dipersepsikan responden berada pada tingkat kelengkapan yang sedang, tidak tergolong sangat lengkap maupun sangat tidak lengkap.

Pada variabel kemudahan pemahaman, skor total yang diperoleh sebesar 129 dengan rata-rata 2.87, sedikit lebih tinggi dibandingkan ketiga variabel lainnya, dengan interpretasi "Cukup Mudah Dipahami". Nilai rata-rata yang lebih tinggi pada variabel ini menunjukkan bahwa secara umum tingkat kesulitan pemahaman terhadap jawaban model relatif sedikit lebih besar dibandingkan tingkat ketidaklengkapan jawaban, meskipun perbedaannya tidak signifikan dan keduanya tetap berada pada kategori interpretasi yang sama, yaitu Cukup.

Pada variabel kepuasan terhadap chatbot, skor total tercatat sebesar 126 dengan rata-rata 2.80, dengan interpretasi "Cukup Puas". Nilai ini identik dengan skor total dan rata-rata pada variabel kelengkapan jawaban, sehingga kedua variabel tersebut berada pada kategori interpretasi yang sama, yaitu Cukup.

Pada variabel kepuasan terhadap agen manusia, skor total dan rata-rata yang diperoleh juga sebesar 126 dan 2.80, dengan interpretasi yang sama, yaitu "Cukup Puas". Kesamaan nilai rata-rata antara kepuasan terhadap chatbot dan kepuasan terhadap agen manusia merupakan temuan yang relevan dalam pengujian implementasi model ini, karena menunjukkan bahwa berdasarkan data yang diperoleh, tidak terdapat perbedaan tingkat kepuasan secara antara kedua jenis layanan yang dibandingkan.

Secara keseluruhan, perbandingan keempat variabel pada Tabel 4.12 menunjukkan bahwa nilai rata-rata berkisar antara 2.80 hingga 2.87, dengan seluruh variabel berada pada kategori interpretasi Cukup. Kesetaraan nilai rata-rata antara kepuasan terhadap chatbot dan kepuasan terhadap agen manusia mengindikasikan bahwa performa chatbot belum menunjukkan perbedaan yang berarti dibandingkan agen manusia secara agregat pada pengujian ini. Dengan demikian, hasil implementasi model pada pengujian ini menunjukkan bahwa chatbot berbasis TinyLLaMA dengan fine-tuning LoRA menghasilkan tingkat penerimaan pengguna yang relatif setara dengan layanan agen manusia pada sampel responden yang diuji, dengan kategori kepuasan yang masih berada pada level Cukup.

4.5.2 Analisis Apakah Model Sudah Optimal

Untuk menentukan apakah model sudah optimal pada penelitian ini didasarkan pada dua sumber data, yaitu hasil evaluasi kuantitatif berbasis metrik BLEU-4 pada tahap fine-tuning dan hasil pengujian persepsi pengguna pada Tabel 4.12. Kedua sumber data tersebut digunakan secara bersamaan agar penilaian terhadap model tidak hanya bertumpu pada satu sudut pandang, mengingat metrik evaluasi otomatis dan persepsi pengguna dapat menggambarkan aspek performa model yang berbeda.

Pada tahap pengujian persepsi pengguna, hasil pada Tabel 4.12 menunjukkan bahwa keempat variabel pengukuran, yaitu kelengkapan jawaban, kemudahan pemahaman, kepuasan terhadap chatbot, dan kepuasan terhadap agen manusia, berada pada rentang rata-rata 2.80–2.87 dengan kategori interpretasi Cukup berdasarkan pedoman pada Tabel 4.13. Nilai ini menunjukkan bahwa secara

persepsi pengguna, model berada pada kategori menengah, tidak tergolong Baik atau Sangat Baik, namun juga tidak berada pada kategori Buruk. Temuan yang relevan untuk dibahas adalah kesetaraan nilai rata-rata kepuasan terhadap chatbot dan agen manusia, yang sama-sama berada pada angka 2.80 dengan interpretasi Cukup Puas, sehingga berdasarkan data yang diperoleh tidak terdapat perbedaan tingkat kepuasan secara agregat antara kedua jenis layanan tersebut.

Apabila kedua sumber data tersebut dibandingkan, terdapat perbedaan kecenderungan antara hasil evaluasi BLEU-4 dan hasil persepsi pengguna. Skor BLEU-4 sebesar 14.54 mengindikasikan tingkat kesesuaian leksikal yang masih rendah terhadap teks referensi, sementara hasil pengujian persepsi pengguna menunjukkan kategori Cukup pada seluruh variabel, dengan tingkat kepuasan terhadap chatbot yang setara dengan agen manusia. Perbedaan kecenderungan ini dapat dijelaskan oleh karakteristik metrik BLEU-4 yang hanya mengukur kecocokan kata atau frasa secara permukaan, tanpa mempertimbangkan apakah jawaban yang berbeda secara leksikal dari referensi tetap dapat dipahami dan dinilai relevan oleh pengguna. Dengan kata lain, jawaban model yang memperoleh skor BLEU-4 rendah tidak secara otomatis dipersepsikan tidak lengkap atau tidak memuaskan oleh responden, sebagaimana ditunjukkan oleh hasil pada Tabel 4.12.

Berdasarkan kedua sumber data tersebut, dapat disimpulkan bahwa dari sisi metrik evaluasi BLEU-4, model masih menunjukkan tingkat kesesuaian leksikal yang rendah terhadap referensi, sementara dari sisi kuesioner persepsi pengguna, model menunjukkan hasil yang berada pada kategori Cukup dan setara dengan layanan agen manusia. Kondisi ini menunjukkan bahwa nilai metrik evaluasi

otomatis yang rendah tidak selalu berbanding lurus dengan rendahnya penerimaan pengguna terhadap jawaban model. Dengan demikian, hasil evaluasi BLEU-4 dan hasil kuesioner pada penelitian ini sebaiknya dimaknai sebagai dua gambaran yang saling melengkapi, yaitu model masih memiliki ruang perbaikan dari sisi kesesuaian leksikal terhadap referensi, namun dari sisi kuesioner persepsi pengguna, model telah menunjukkan tingkat penerimaan yang tidak berbeda jauh dari layanan agen manusia pada pengujian ini.

4.6 Integrasi Islam

Hasil pada subbab sebelumnya menunjukkan bahwa di antara tiga skenario konfigurasi LoRA Rank yang diujikan, Skenario 3 ($r=32$, $\alpha=64$) mencatat skor BLEU-4 tertinggi sebesar 14.54, namun selisihnya dengan Skenario 2 ($r=16$, $\alpha=32$) yang mencapai 14.49 hanya sebesar 0.05 poin. Hasil ini menunjukkan bahwa penambahan kapasitas rank di atas $r=16$ tidak memberikan peningkatan kualitas keluaran yang signifikan, sehingga yang menentukan keberhasilan suatu konfigurasi bukanlah besarnya rank yang digunakan, melainkan sejauh mana hasil yang dicapai benar-benar optimal. Dalam perspektif Islam, prinsip ini sejalan dengan firman Allah Subhanahu wa Ta'ala dalam QS. Al-Kahf [18]: 7:

إِنَّا جَعَلْنَا مَا عَلَى الْأَرْضِ زِينَةً لِّهَا لِنَبْلُوَهُمْ أَيُّهُمْ أَحْسَنُ عَمَلًا (٦)

"Sesungguhnya Kami telah menjadikan apa yang ada di atas bumi sebagai perhiasan baginya agar Kami menguji mereka siapakah di antaranya yang lebih baik perbuatannya." (QS. Al-Kahf [18]: 7)

Ayat diatas menjelaskan bahwa Allah menjadikan beraneka macam hewan, tumbuh-tumbuhan, dan kekayaan alam sebagai perhiasan bumi untuk menguji

manusia dalam menyikapi keindahan tersebut, sehingga Allah mengetahui secara nyata siapa di antara mereka yang terbaik perbuatannya bukan siapa yang terbanyak atau terbesar capaiannya (Kementerian Agama Republik Indonesia, n.d.-b). Tafsir Tahlili menambahkan bahwa ujian ini berkaitan dengan bagaimana manusia memanfaatkan sumber daya yang tersedia, apakah digunakan untuk kemaslahatan atau sebaliknya. Penekanan pada kata *aḥsanu 'amalā* (perbuatan yang terbaik) inilah yang relevan dengan temuan penelitian ini: penambahan rank dari $r=16$ menjadi $r=32$ pada dasarnya adalah penambahan sumber daya komputasi, namun tidak menghasilkan amal/kerja yang lebih baik secara substansial. Hal ini menjadi pengingat bahwa nilai sebuah ikhtiar tidak diukur dari skala atau besarnya sumber daya yang dikerahkan, melainkan dari kualitas dan ketepatan hasil yang dicapai sesuai dengan kebutuhan. Skenario 3 dipilih sebagai model terbaik bukan karena memiliki parameter terbesar, tetapi karena pada batas itulah keseimbangan antara efisiensi dan kualitas keluaran tercapai.

Selanjutnya, hasil uji implementasi terhadap 45 responden menunjukkan adanya kesenjangan antara evaluasi metrik otomatis dan persepsi pengguna secara langsung. Meskipun skor BLEU-4 yang diperoleh tergolong rendah hingga moderat, hasil kuesioner pada Tabel 4.12 menunjukkan bahwa seluruh variabel pengukuran berada pada kategori Cukup dengan rata-rata berkisar antara 2.80 hingga 2.87, dengan nilai kepuasan terhadap chatbot dan agen manusia yang sama-sama mencapai angka 2.80. Hal ini mengindikasikan bahwa keterbatasan leksikal model tidak serta-merta berimplikasi pada rendahnya penerimaan pengguna, namun kategori Cukup yang dicapai juga mencerminkan bahwa chatbot berbasis

TinyLLaMA masih memiliki keterbatasan inheren dalam menangani pertanyaan yang membutuhkan penilaian kontekstual mendalam. Dalam perspektif Islam, kenyataan ini dapat dipahami melalui firman Allah Subhanahu wa Ta'ala dalam QS. Al-Mu'minūn [23]: 62:

وَلَا تُكَلِّفُ نَفْسًا إِلَّا وُسْعَهَا وَلَدَيْنَا كِتَابٌ يَنْطِقُ بِالْحَقِّ وَهُمْ لَا يُظْلَمُونَ (٦٢)

"Kami tidak membebani seorang pun, kecuali menurut kesanggupannya. Pada Kami ada suatu catatan yang menuturkan dengan sebenarnya dan mereka tidak dizalimi." (QS. Al-Mu'minūn [23]: 62)

Ayat di atas menerangkan bahwa Allah tidak membebani seseorang dengan amalan melebihi kesanggupannya, sehingga tidak sewajarnya sesuatu dinilai gagal hanya karena tidak mencapai standar yang berada di luar batas kemampuannya (Kementerian Agama Republik Indonesia, n.d.-c). Tafsir Tahlili memperkuat hal ini dengan menjelaskan bahwa "tidak ada syariat yang diwajibkan yang berat dilaksanakan oleh hambanya dan di luar batas kemampuannya" (Kementerian Agama Republik Indonesia, n.d.-c); suatu kewajiban hanya akan terasa berat bagi yang memandangnya melebihi kesanggupan, bukan karena kewajiban itu sendiri tidak proporsional. Prinsip wus'ahā (sesuai kesanggupan) ini relevan untuk memaknai posisi chatbot dalam penelitian ini: TinyLLaMA-1.1B adalah model bahasa berkapasitas kecil yang dilatih pada dataset sintetis terbatas, sehingga kemampuannya menjawab pertanyaan properti memang berada pada batas kesanggupannya bukan kapasitas tanpa batas seperti agen manusia yang memiliki penilaian kontekstual dan pengalaman lapangan. Bahwa nilai kepuasan terhadap chatbot dan agen manusia sama-sama berada pada kategori Cukup bukanlah bukti

kegagalan mutlak chatbot, melainkan cerminan wajar dari prinsip bahwa setiap entitas termasuk sistem kecerdasan buatan memiliki batas kesanggupan yang harus dimaklumi, sebagaimana ditegaskan pula pada lanjutan ayat ini bahwa "mereka tidak dizalimi" (wa hum lā yuẓlamūn). Artinya, keberadaan agen manusia sebagai layanan komplementer justru menjamin bahwa pengguna tidak dirugikan ketika kebutuhan mereka berada di luar batas kesanggupan chatbot, sehingga kedua sistem chatbot dan agen manusia pada hakikatnya saling melengkapi sesuai dengan kapasitas masing-masing.

Secara keseluruhan, kedua ayat ini menegaskan bahwa keberhasilan pengembangan chatbot konsultan properti berbasis TinyLLaMA dalam penelitian ini tidak semestinya dinilai dari besarnya sumber daya yang dikerahkan maupun dari kesempurnaan metrik tekstual semata, melainkan dari sejauh mana sistem tersebut bekerja secara optimal sesuai dengan kapasitas dan kesanggupannya. Dengan demikian, performa yang berada pada kategori Cukup pada seluruh variabel pengukuran dapat dimaknai sebagai cerminan dari prinsip wus'ahā, yaitu hasil yang sesuai dengan batas kesanggupan model pada konfigurasi dan dataset yang digunakan saat ini, sekaligus menjadi dasar bagi ruang penyempurnaan pada penelitian selanjutnya.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian, fine-tuning TinyLLaMA-1.1B-Chat-v1.0 menggunakan metode LoRA mampu mengadaptasi model umum menjadi chatbot konsultan properti berbahasa Indonesia, meskipun belum mencapai hasil yang optimal. Dataset sintetis sebanyak 3024 pasangan instruction–output dengan tujuh kategori intent.

Dari sisi variasi LoRA Rank memberikan pengaruh terhadap kualitas keluaran model, dengan Skenario 3 ($r=32$, $\alpha=64$) mencatat skor BLEU-4 tertinggi sebesar 14.54, diikuti Skenario 2 ($r=16$, $\alpha=32$) sebesar 14.49 dan Skenario 1 ($r=8$, $\alpha=16$) sebesar 12.76. Selisih yang sangat kecil antara Skenario 2 dan Skenario 3 (0.05 poin) menunjukkan bahwa peningkatan rank di atas $r=16$ tidak menghasilkan perbedaan kesesuaian tekstual yang substansial, sehingga Skenario 3 dipilih sebagai model terbaik dan diimplementasikan pada sistem chatbot berbasis Laravel. Skor BLEU-4 pada ketiga skenario yang berada pada rentang yang rendah yaitu berkisar 12.76–14.54 yang mencerminkan keterbatasan kesesuaian leksikal akibat karakteristik metrik BLEU-4 itu sendiri serta dominasi korpus bahasa Inggris pada model dasar TinyLLaMA-1.1B dan juga keseragaman pola kalimat pada dataset sintetis.

Dari sisi persepsi pengguna, hasil uji implementasi terhadap 45 responden menunjukkan bahwa keempat variabel pengukuran kelengkapan jawaban,

kemudahan pemahaman, kepuasan terhadap chatbot, dan kepuasan terhadap agen manusia berada pada kategori Cukup dengan rata-rata 2.80–2.87, dengan kepuasan terhadap chatbot dan agen manusia yang setara (2.80). Hal ini menunjukkan tidak terdapat perbedaan tingkat kepuasan secara signifikan antara kedua jenis layanan tersebut, sekaligus mengindikasikan adanya kesenjangan antara evaluasi metrik otomatis dan persepsi pengguna secara langsung.

Dengan demikian, hasil evaluasi BLEU-4 dan kuesioner pada penelitian ini mengukur aspek yang berbeda namun mengarah pada simpulan yang sejalan. Model masih memiliki ruang perbaikan dari sisi kesesuaian leksikal, namun dari sisi persepsi pengguna telah menunjukkan tingkat penerimaan yang tidak berbeda jauh dari layanan agen manusia. Secara keseluruhan, fine-tuning LoRA mampu mengadaptasi TinyLLaMA ke domain properti Indonesia, namun belum mencapai hasil yang optimal mengingat skor BLEU-4 yang masih rendah hingga moderat serta tingkat kepuasan pengguna yang hanya berada pada kategori Cukup.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat dijadikan pertimbangan untuk pengembangan sistem maupun penelitian selanjutnya.

5.2.1 Saran Pengembangan Sistem

Pengembangan sistem chatbot konsultan properti sebaiknya difokuskan pada peningkatan kualitas dan kedalaman jawaban yang dihasilkan model. Hal ini dapat dilakukan dengan memperkaya variasi dataset pelatihan, khususnya dengan

menambahkan skenario percakapan yang lebih kompleks dan beragam, sehingga model memiliki kemampuan generalisasi yang lebih baik terhadap pertanyaan di luar pola yang telah dilatih. Selain itu, penerapan pendekatan hybrid dapat dipertimbangkan sebagai langkah lanjutan, di mana chatbot menangani pertanyaan yang bersifat rutin dan terstruktur, sementara pertanyaan yang memerlukan pertimbangan lebih kontekstual tetap diarahkan kepada agen manusia. Pendekatan ini relevan mengingat hasil penelitian menunjukkan bahwa tingkat kepuasan pengguna terhadap chatbot maupun agen manusia sama-sama berada pada kategori Cukup, yang mengindikasikan masih adanya ruang peningkatan pada kemampuan model dalam menangani pertanyaan yang memerlukan penilaian kontekstual mendalam.

5.2.2 Saran Penelitian Selanjutnya

Penelitian selanjutnya disarankan mempertimbangkan beberapa perbaikan mendasar, baik dari sisi metrik evaluasi, pemilihan model, maupun kesesuaian metodologi dengan karakteristik tugas yang dikerjakan.

1. Dari sisi metrik evaluasi, penelitian ini hanya menggunakan BLEU-4 yang memiliki keterbatasan signifikan untuk tugas percakapan konsultatif karena hanya mengukur kecocokan n-gram leksikal tanpa mempertimbangkan relevansi semantik. Keterbatasan ini tercermin dari kesenjangan antara skor BLEU-4 yang rendah dan penerimaan pengguna yang cukup positif pada penelitian ini. Penelitian selanjutnya disarankan mengadopsi metrik berbasis representasi semantik seperti BERTScore, ROUGE, atau MoverScore untuk memperoleh gambaran evaluasi yang lebih komprehensif dan sesuai dengan

karakteristik domain percakapan.

2. Dari sisi kesesuaian model dasar, TinyLLaMA-1.1B yang berbasis bahasa Inggris memiliki keterbatasan dalam menangani nuansa linguistik, istilah lokal, dan gaya percakapan formal-informal bahasa Indonesia. Penelitian selanjutnya disarankan mengeksplorasi model yang secara eksplisit dilatih pada korpus bahasa Indonesia, seperti IndoBERT, IndoGPT, atau Merak, sebagai fondasi sebelum dilakukan fine-tuning domain spesifik.
3. Dari sisi dataset, ketergantungan penuh pada data sintetis dari GPT berpotensi menghasilkan distribusi pola kalimat yang seragam sehingga model kurang mampu menggeneralisasi terhadap pertanyaan nyata pengguna. Penelitian selanjutnya disarankan menggabungkan data sintetis dengan data percakapan nyata dari interaksi agen properti dengan calon pembeli guna meningkatkan variasi linguistik yang lebih realistis.
4. Dari sisi konfigurasi LoRA, penelitian ini hanya menargetkan query dan value projection. Eksplorasi terhadap layer tambahan seperti key projection, output projection, atau feed-forward layer, serta integrasi dengan pendekatan Retrieval-Augmented Generation (RAG), dapat dipertimbangkan untuk meningkatkan cakupan adaptasi dan akurasi jawaban faktual terkait regulasi dan prosedur properti yang bersifat dinamis.
5. Dari sisi evaluasi pengguna, instrumen kuesioner Likert yang digunakan belum disertai panduan interaksi terstandar sehingga pengalaman antara responden berpotensi tidak sebanding. Penelitian selanjutnya disarankan merancang protokol pengujian yang lebih terstruktur dengan skenario

pertanyaan seragam, dilengkapi wawancara kualitatif untuk memperoleh pemahaman lebih mendalam mengenai kebutuhan dan ekspektasi pengguna.

DAFTAR PUSTAKA

- Akhsya, R. M., Hakiem, H., & Nawawi, M. K. (2022). Faktor-faktor yang mempengaruhi keputusan konsumen dalam pembelian properti syariah. *Jurnal Pendidikan dan Konseling (JPDK)*, 4(5), 3827–3839. <https://doi.org/10.31004/jpdk.v4i5.7219>
- Al Jupri, M. F. U., & Fasa, M. I. (2025). Pengaruh penggunaan teknologi chatbot dalam meningkatkan layanan pelanggan pada platform e-commerce: Studi kasus pada Shopee Indonesia. *Jurnal Media Akademik (JMA)*, 3(3). <https://doi.org/10.62281>
- Basilico, M. (2024). Politecnico di Torino Design, Implementation and Evaluation of a Chatbot for Accounting Firm: A Fine-Tuning Approach with Two Novel Dataset.
- Biderman, D., Portes, J., Gonzalez Ortiz, J. J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., & Cunningham, J. P. (2024). LoRA learns less and forgets less. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=aloEru2qCG>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *arXiv*. <https://doi.org/10.48550/arXiv.2305.14314>
- Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., & Joty, S. (2024). Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv*. <https://doi.org/10.48550/arXiv.2403.02990>
- Faturahman, M. A., Studi, P., & Informasi, S. (2023). Perancangan Aplikasi Sistem Informasi Agen Real Estate Berbasis Website Di Perusahaan Broker Perumahan Pt Berkas Anugerah Prima Kota Bandung. 1–15. <https://doi.org/10.32897/infotronik.202x.6.2.1566>
- Firman, A. (2022). Pengaruh bauran pemasaran terhadap keputusan konsumen dalam membeli rumah. *JESYA: Jurnal Ekonomi & Ekonomi Syariah*, 5(2). <https://doi.org/10.36778/jesya.v5i2.773>
- Fu, X.-Y., Laskar, M. T. R., Khasanova, E., Chen, C., & Bhushan, S. TN. (2024). Tiny Titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *arXiv*. <https://doi.org/10.48550/arXiv.2402.00841>
- Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2024). More Than a Bot? The Impact of Disclosing Human Involvement on Customer Interactions with Hybrid Service Agents More Than a Bot? The Impact of Disclosing Human Involvement on Customer Interactions with Hybrid Service

Agents. September 2025. <https://doi.org/10.1287/isre.2022.0152>

Han, D., Song, X., & Cui, Y. (2020). An extractive chat summary generation method for ecommerce chatbots. *DEStech Transactions on Computer Science and Engineering*. <https://doi.org/10.12783/dtcse/cisnr2020/35169>

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>

Hu, E. J., 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*. <https://doi.org/10.48550/arXiv.2106.09685>

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

Kementerian Agama Republik Indonesia. (n.d.-a). Al-Qur'an Surah Al-Hujurat (49): Ayat 1-18. Qur'an Kemenag. <https://quran.kemenag.go.id/quran/per-ayat/surah/49?from=1&to=18>

Kementerian Agama Republik Indonesia. (n.d.-b). Al-Qur'an Surah Al-Kahf (18): Ayat 1-110. Qur'an Kemenag. <https://quran.kemenag.go.id/quran/per-ayat/surah/18?from=1&to=110>

Kementerian Agama Republik Indonesia. (n.d.-c). Al-Qur'an Surah Al-Mu'minun (23): Ayat 1-118. Qur'an Kemenag. <https://quran.kemenag.go.id/quran/per-ayat/surah/23?from=1&to=118>

Kementerian Agama Republik Indonesia. (n.d.-d). Al-Qur'an Surah Az-Zumar (39): Ayat 1-75. Qur'an Kemenag. <https://quran.kemenag.go.id/quran/per-ayat/surah/39?from=1&to=75>

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. (2025). RoRA: Efficient fine-tuning of LLM with reliability optimization for rank adaptation. *arXiv*. <https://doi.org/10.48550/arXiv.2501.04315>

Maheshwari, G., Ivanov, D., & El Haddad, K. (2024). Efficacy of synthetic data as a benchmark. *arXiv*. <https://doi.org/10.48550/arXiv.2409.11968>

Manik, A., Siahaan, O. T. R., & Siregar, J. K. G. (2026). Analisis pandangan dan tanggapan pengguna terhadap chatbot berbasis AI: Studi kasus ChatGPT. *ICo SCIENCE: Journal of Intelligent Computing and Advanced Data Science*, 1(1).

Meyer, S., Singh, S., Tam, B., Ton, C., & Ren, A. (2024). A comparison of LLM finetuning methods & evaluation metrics with travel chatbot use case. *arXiv*. <https://doi.org/10.48550/arXiv.2408.03562>

Pandya, K., & Holia, M. (2023). Automating customer service using LangChain:

- Building custom open-source GPT chatbot for organizations. arXiv. <https://doi.org/10.48550/arXiv.2310.05421>
- Paula, R., Aguiar Neto, D., Romero, D., & Guerra, P. (2021). Evaluation of synthetic datasets generation for intent classification tasks in Portuguese. In E. S. Ruiz & T. T. Torrent (Eds.), *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology* (pp. 265–274). Association for Computational Linguistics. <https://aclanthology.org/2021.stil-1.29/>
- 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. *JURNAL ILMIAH INFORMATIKA*, 12(01), 31–36. <https://doi.org/10.33884/jif.v12i01.8189>
- Septiana, C., Shinta, A., & Rohyati, E. (2021). Proses Pengambilan Keputusan pada Konsumen untuk Membeli Rumah dari PT X Berdasarkan Alasan Pro-Lingkungan Hidup. *Jurnal Psikologi*, 17(1), 48–59. Retrieved from <https://ejournal.up45.ac.id/index.php/psikologi/article/view/818>
- Soetiyono, A., Kurnia, Y., & Kurnia, R. (2024). Pengaruh Penggunaan Chatbot dan Asisten Virtual terhadap Peningkatkan Kepuasan Pelanggan serta Dampaknya terhadap Pengambilan Keputusan Pembelian. *ECo-Buss*, 6(3), 1367–1381. <https://doi.org/10.32877/eb.v6i3.1169>
- Titoni, E., Firgana, D., Aditya, B., Pribadi, T. A., & Lianto, R. (2024). Peran chatbot artificial intelligence dalam membentuk kepuasan pelanggan. *Technologia: Jurnal Ilmiah*, 15(1), 20. <https://doi.org/10.31602/tji.v15i1.13220>
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. arXiv. <https://doi.org/10.48550/arXiv.2212.10560>
- Wu, J., Zhu, Y., Shen, L., & Lu, X. (2024). GEB-1.3B: Open lightweight large language model. arXiv. <https://doi.org/10.48550/arXiv.2406.09900>
- Xia, M., Gao, T., Zeng, Z., & Chen, D. (2024). Sheared LLaMA: Accelerating language model pre-training via structured pruning. arXiv. <https://doi.org/10.48550/arXiv.2310.06694>
- Yang, A., Liu, K., Liu, J., Lyu, Y., & Li, S. (2018). Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. arXiv. <https://doi.org/10.48550/arXiv.1806.03578>
- Yang, Y., Tao, C., & Fan, X. (2025). LoRA-LiteE: A computationally efficient framework for chatbot preference-tuning. arXiv. <https://doi.org/10.48550/arXiv.2411.09947>

- Ye, Z., Li, D., Hu, Z., Lan, T., Sha, J., Zhang, S., Duan, L., Zuo, J., Lu, H., Zhou, Y., & Tang, M. (2024). mLoRA: Fine-tuning LoRA adapters via highly-efficient pipeline parallelism in multiple GPUs. arXiv. <https://doi.org/10.48550/arXiv.2312.02515>
- Yu, H., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *Journal Of Medical Artificial Intelligence*, 7. doi:10.21037/jmai-23-136
- Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv. <https://doi.org/10.48550/arXiv.2401.02385>
- 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>