BAB 1 PENDAHULUAN

1.1. Latar Belakang

Layanan pelanggan (*customer service*) memiliki peran krusial dalam menjaga kepuasan pengguna pada berbagai sektor, termasuk sektor pendidikan. Di lingkungan kampus, mahasiswa sebagai "pelanggan" membutuhkan akses informasi akademik maupun administratif secara cepat dan akurat. Namun, metode layanan pelanggan konvensional seperti telepon, email, atau loket fisik masih menghadapi beberapa hambatan. Contohnya, waktu tunggu yang lama dapat menyebabkan ketidakpuasan, serta keterbatasan kemampuan untuk menangani permasalahan yang lebih kompleks [1]. Kondisi ini memicu pencarian solusi layanan pelanggan yang lebih efektif dan efisien untuk meningkatkan kualitas pelayanan di kampus.

Saat ini, kemajuan teknologi *Large Language Models* (LLM) menawarkan cara baru untuk mengembangkan *chatbot* interaktif. LLM mampu memahami dan menghasilkan bahasa alami secara kontekstual, sehingga cocok untuk interaksi berbasis dialog [2]. Berbagai penelitian telah memanfaatkan LLM untuk membangun *chatbot* cerdas di berbagai *domain*, termasuk pelayanan kampus [3]. Meskipun demikian, proses *fine-tuning* LLM sering memerlukan sumber daya komputasi yang besar, sehingga membatasi pengembangannya di lingkungan yang memiliki keterbatasan perangkat keras [4].

Untuk mengatasi kendala tersebut, teknik *Quantized Low-Rank Adaptation* (QLoRA) diperkenalkan sebagai metode *fine-tuning* yang lebih efisien. QLoRA menerapkan kuantisasi presisi 4-bit pada model, yang memungkinkan penyesuaian model berukuran besar dengan konsumsi memori yang lebih rendah tanpa mengorbankan kinerja [4]. Pendekatan ini juga berakar pada metode *Low-Rank Adaptation* (LoRA), yang telah terbukti efektif dalam mengurangi kompleksitas komputasi sambil tetap menjaga

kualitas hasil [5]. Dengan demikian, QLoRA membuka peluang penerapan LLM di berbagai institusi yang memiliki keterbatasan infrastruktur.

Meskipun sejumlah penelitian telah membuktikan keunggulan *chatbot* berbasis LLM dalam mendukung layanan pelanggan [6][7], penelitian khusus yang mengevaluasi pemanfaatan QLoRA pada layanan kampus masih terbatas. Penerapan QLoRA berpotensi menjadi solusi yang tepat untuk membangun *chatbot* yang efisien dan relevan di lingkungan komputasi yang terbatas dalam memberikan respons seperti jadwal kuliah, prosedur akademik, atau pelayanan administratif sehari-hari. Oleh karena itu, pengembangan *chatbot* kampus berbasis QLoRA menjadi topik yang menarik untuk diteliti lebih lanjut, mengingat dampaknya yang signifikan terhadap kepuasan mahasiswa dan efisiensi operasional di lingkungan pendidikan.

1.2. Rumusan Masalah

Penelitian ini difokuskan untuk menyelesaikan isu-isu utama terkait pengembangan *chatbot* layanan pelanggan di lingkungan kampus menggunakan teknik QLoRA. Rumusan masalah yang diidentifikasi dalam penelitian ini adalah sebagai berikut:

1. Bagaimana proses *fine-tuning* untuk mengoptimalkan *chatbot* berbasis LLM menggunakan teknik QLoRA dalam pemenuhan kebutuhan layanan pelanggan di lingkungan kampus?
2. Bagaimana performa *chatbot* yang telah di-*fine-tuning* menggunakan QLoRA?

1.3. Tujuan dan Manfaat

Berdasarkan rumusan masalah di atas, tujuan utama dari penelitian ini adalah untuk mengoptimalkan dan mengevaluasi *chatbot customer service* kampus berbasis LLM melalui teknik QLoRA. Secara rinci, tujuan penelitian ini adalah sebagai berikut:

1. Mengoptimalkan proses *fine-tuning chatbot* berbasis LLM dengan menggunakan teknik QLoRA dalam pemenuhan kebutuhan layanan pelanggan di lingkungan kampus.

2. Mengevaluasi performa *chatbot* yang telah di-*fine-tuning* dengan teknik QLoRA.

Manfaat yang diharapkan dari penelitian ini adalah peningkatan efisiensi dan produktivitas layanan pelanggan di lingkungan kampus melalui penggunaan sistem *chatbot* yang responsif dan relevan. Di samping itu, penelitian ini juga diharapkan dapat memberikan kontribusi teknis dalam pengembangan *chatbot* berbasis LLM, khususnya dalam pemanfaatan teknik QLoRA yang efisien dalam penggunaan sumber daya komputasi.

1.4. Batasan Masalah

Agar penelitian ini dapat dilaksanakan secara efektif dan efisien serta dengan sumber daya yang tersedia, maka penelitian ini dibatasi pada beberapa aspek sebagai berikut:

1. Dataset: Penelitian ini menggunakan dataset FAQ *Helpdesk* Mahasiswa yang terdiri dari 379 baris data, mencakup pertanyaan, konteks, dan respons terkait layanan kampus, dengan fokus pada pertanyaan umum seputar administrasi dan akademik.
2. Metrik Evaluasi: Evaluasi performa *chatbot* dilakukan menggunakan metrik kuantitatif dan kualitatif sebagai berikut:
    - *Cross-Entropy Loss*: Metrik ini digunakan untuk mengukur performa model selama pelatihan, dengan menghitung perbedaan antara distribusi probabilitas prediksi model dan distribusi referensi.
    - BLEU (*Bilingual Evaluation Understudy*): Metrik ini mengukur kesesuaian *n-gram* antara respons model dengan jawaban referensi, memberikan gambaran tentang akurasi linguistik respons model.
    - ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*): Digunakan untuk mengevaluasi relevansi dan cakupan informasi dalam respons model dengan fokus pada *recall* dan kesesuaian frasa.

- Evaluasi Manual: Selain metrik kuantitatif, dilakukan evaluasi manual terhadap respons model untuk menilai relevansi, koherensi, dan kualitas linguistik pada konteks tertentu.

3. Model yang Digunakan**:** Penelitian ini hanya menguji teknik QLoRA pada model Mistral-7B tanpa membandingkannya dengan teknik *fine-tuning* lainnya atau model LLM lainnya.

1.5. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental untuk mengeksplorasi dan mengoptimalkan pengembangan *chatbot* layanan pelanggan berbasis LLM dengan teknik QLoRA di lingkungan kampus. Beberapa tahapan utama dalam metode penelitian ini meliputi:

1. Studi Literatur: Peneliti melakukan kajian pustaka untuk memahami berbagai konsep yang terkait dengan *chatbot*, LLM, serta teknik *fine-tuning*, terutama yang menggunakan QLoRA.
2. Pengumpulan Data: yang digunakan dalam penelitian ini adalah FAQ *Helpdesk* Mahasiswa yang mencakup berbagai pertanyaan umum tentang administrasi dan akademik di lingkungan kampus.
3. Pengembangan Model *Chatbot*: Proses pengembangan dimulai dengan pemilihan model LLM yang sesuai, yaitu Mistral-7B. Model ini kemudian di-*fine-tune* menggunakan teknik QLoRA dengan dataset yang telah disiapkan.
4. Pengujian dan Evaluasi: Setelah model selesai dilatih, dilakukan serangkaian pengujian dan evaluasi untuk mengukur performa *chatbot* menggunakan metrik kuantitatif (*Cross-Entropy Loss*, BLEU, ROUGE) serta evaluasi manual untuk menilai kualitas dan relevansi respons yang dihasilkan.

Melalui metode penelitian ini, diharapkan dapat dikembangkan *chatbot* layanan pelanggan yang efisien dan efektif dengan menggunakan teknik QLoRA untuk meminimalkan penggunaan sumber daya komputasi tanpa mengorbankan kualitas respons.