202210370311251
Muhammad Hauzan Afif
Teknik Informatika

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Penelitian ini diinisiasi karena adanya kerja sama antara Prodi Informatika Universitas Muhammadiyah Malang dan PT Bima Sakti. Setelah terbentuknya kerja sama antara dua kelompok, maka dibentuklah sebuah tim kecil beranggotakan 5 orang untuk menyelesaikan sebuah masalah yang ada pada industri nyata. Masalah yang ditemukan adalah bagaimana caranya mahasiswa atau staf mahasiswa dalam mengatasi akses informasi mahasiswa. Setelah melakukan analisa, kami menemukan sebuah masalah tentang efisiensi akses informasi akademik mahasiswa yang dirasa belum optimal jika dibandingkan dengan perkembangan teknologi saat ini.

Kami sepakat dan mengusulkan solusi dari permasalahan ini dengan adanya penerapan atau pengembangan asisten virtual berbasis AI Generatif dengan menggunakan lingkungan 3D Unity Engine, dikarenakan bukan hanya bisa memberikan jawaban secara *real-time,* pengguna juga bisa berinteraksi secara langsung dengan visual yang telah disajikan.

Pada pengembangan sistem ini, penulis bertanggung jawab dan fokus pada sisi backend atau sistem utama dalam pembuatan asisten virtual berbasis AI generatif. Tugas utama penulis berfokus pada proses *fine-tuning* model AI sebagai otak asisten virtual, pembangunan REST API untuk menjembatani komunikasi data *frontend* dan *backend.* Penulis juga memiliki tugas untuk memastikan logika kecerdasan buatan mampu memahami konteks akademik UMM sebelum di *deploy* atau didistribusikan ke antarmuka Unity yang dikerjakan oleh rekan tim.

Model yang digunakan oleh penulis dalam penelitian ini adalah Llama 3.2 3B. Dalam pemilihan model penulis memilah semua varian model berdasarkan efisiensi pada komputasi, dikarenakan tahapan akhir dari pengembangan asisten virtual generatif AI ini berjalan pada perangkat lokal. Walaupun berjalan pada perangkat

lokal tanpa adanya server, model harus tetap memiliki kapabilitas penalaran yang memadai. Menurut penulis, model varian Llama 3.2 3B maupun tergolong model dengan parameter yang relatif kecil, model ini memiliki kapabilitas yang sangat mumpuni untuk melakukan pelatihan *fine-tuning* menggunakan metode pelatihan Parameter-Efficient Fine-Tuning (PEFT) dengan teknik LoRA. Teknik tersebut yang akan digunakan oleh penulis karena efisiensi pada saat melatih model, karena hanya sebagian kecil bobotnya saja yang dipakai, sehingga saat proses adaptasi domain dapat dilakukan tanpa memerlukan infrastruktur server berspesifikasi tinggi.

Untuk penerapan model lokal penulis memilih memakai Ollama, karena sistem Ollama memberikan dukungan lokal atau bisa disebut localhost, lalu sistem tersebut memudahkan untuk melakukan integrasi. Dengan adanya Arsitektur ini memungkinkan aplikasi Unity dalam mengirimkan prompt dan menerima balasan secara langsung melalui skrip C#, menjadikan integrasi antara visual 3D dan logika AI lebih stabil dan responsif.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah ditulis, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana caranya mengolah data mentah akademik menjadi dataset dengan format instruksi yang siap digunakan untuk melatih model?
2. Bagaimana penerapan metode *fine-tuning* LoRA pada model Llama 3.2, khususnya dalam pengaturan *hyperparameter* (seperti *rank* dan *alpha*), agar model mampu memahami konteks akademik UMM?
3. Bagaimana mekanisme integrasi model yang sudah dilatih ke dalam server Ollama agar dapat terhubung dan mengirimkan data ke aplikasi Unity melalui REST API?
4. Bagaimana performa model setelah dilakukan *fine-tuning* jika dibandingkan dengan model aslinya, dilihat dari metrik evaluasi (seperti Perplexity atau ROUGE) dan relevansi jawabannya?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang ditulis, tujuan dari penelitian ini adalah:

1. Menyusun dataset informasi akademik ke dalam format instruksi agar sesuai untuk proses pelatihan model.
2. Menerapkan metode *fine-tuning* menggunakan teknik LoRA pada model Llama 3.2 agar model memiliki pengetahuan tentang konteks akademik UMM.
3. Membangun sistem integrasi antara *backend* LLM dengan antarmuka Unity untuk memungkinkan komunikasi data.
4. Mengukur kinerja model hasil *fine-tuning* dan membandingkannya dengan model dasar menggunakan metrik evaluasi yang relevan.

1.4 Manfaat Penelitian

Penulis berharap dengan adanya penelitian ini dapat memberikan manfaat bagi berbagai pihak, antara lain:

1. **Bagi Mahasiswa**: Dapat memudahkan mahasiswa dalam mengakses informasi secara mandiri dan cepat melalui tanya jawab interaktif, tanpa harus menunggu jam operasional layanan.
2. **Bagi Program Studi**: Dapat membantu efisiensi layanan akademik dengan adanya sistem otomatis untuk menjawab pertanyaan-pertanyaan umum yang sering ditanyakan berulang kali atau pertanyaan *generic,* sehingga dapat membuat staf fokus pada masalah yang lebih kompleks. .
3. **Bagi Universitas Muhammadiyah Malang**: Menjadikan salah satu referensi penerapan teknologi AI generatif yang terjadi di lingkungan kampus sebagai upaya mendukung inovasi digitalisasi layanan.
4. **Bagi Mitra Industri (PT Bima Sakti)**: Memberikan contoh nyata tentang potensi dan performa dari penggunaan model open-source yang telah melalui prose fine-tuning untuk kebutuhan secara spesifik dengan domain yang ditentukan. Hasil penelitian ini juga bisa menjadi bahan pertimbangan teknis bagi perusahan dalam pengembangan produk berbasis AI di masa depan jika diperlukan.

5. **Bagi Pengembangan Ilmu Pengetahuan**: Menambahkan referensi teknis terkait penerapan metode *fine-tuning* (LoRA) pada LLM dan proses integrasinya ke aplikasi *client-server* menggunakan Ollama dan Unity. Hal ini dapat menjadi acuan bagi mahasiswa atau pengembang lain yang ingin membangun sistem yang serupa dengan sumber daya komputasi yang efisien.

1.5 Batasan Penelitian

Agar mencegah pembahasan tidak meluas dan tetap terarah, penelitian ini dibatasi pada ruang lingkup sebagai berikut:

1. **Model Dasar:** Penelitian ini menggunakan model *Llama 3.2* sebagai base model atau basis pengembangan.
2. **Metode Pelatihan:** Teknik *fine-tuning* yang akan diterapkan fokus pada metode *Low-Rank Adaptation* (LoRA). Penelitian ini tidak melakukan perbandingan performa dengan metode *fine-tuning* lainnya maupun pelatihan ulang model dasar yang dilakukan secara penuh (*full fine-tuning*).
3. **Perangkat Lunak:** Proses pelatihan model dilakukan menggunakan pustaka Python (seperti *transformers* dan *peft*), sedangkan untuk menjalankan model di sisi server (*deployment*) akan menggunakan bantuan *Ollama* di perangkat lokal.
4. **Data:** Dataset yang digunakan terbatas pada informasi akademik Program Studi Informatika UMM yang bersifat publik dan tidak mengandung data privasi mahasiswa yang sensitif.
5. **Fungsionalitas Sistem:** Sistem ini dibangun hanya berfungsi sebagai pemberi informasi dengan cara tanya-jawab langsung dengan pengguna. Sistem tidak memiliki hak akses untuk mengubah, menambah, atau memanipulasi data nilai maupun administrasi yang terjadi pada sistem akademik pusat (SIAKAD).
6. **Lingkungan Implementasi:** Pengembangan dan pengujian sistem dilakukan dalam lingkungan jaringan lokal (*localhost*), di mana server API dan aplikasi *client* (Unity) berjalan pada jaringan yang sama.

7. **Antarmuka:** Sisi antarmuka pengguna menggunakan *Unity Engine*, namun fokus penelitian penulis terbatas pada penyediaan layanan *backend* dan API untuk antarmuka tersebut.