BAB I
PENDAHULUAN

1.1. Latar Belakang

Platform Roblox telah bertransformasi menjadi ekosistem metaverse terbesar di dunia, dibuktikan dengan basis pengguna aktif harian yang tumbuh pesat dari 85,3 juta pengguna aktif harian pada akhir 2024 menjadi puncaknya 151,5 juta pada kuartal ketiga 2025 [1], [2]. Fenomena ini menjadi indikasi nyata bahwa isu keamanan anak dan remaja di platform ini bukan sekadar wacana akademis, melainkan risiko operasional dan hukum yang serius bagi penyedia platform itu sendiri. Di Indonesia, data survei APJII tahun 2025 mencatat sekitar 2,07% pengguna internet aktif bermain Roblox [3]. Sayangnya, popularitas dan kebebasan berekspresi di platform ini membawa kerentanan serius terhadap tata krama dan keamanan ruang digital.

Demografi pengguna platform ini memegang peran penting dalam memicu kerentanan tersebut. Skala pengguna sebesar ini seharusnya berbanding lurus dengan tanggung jawab keamanan, terutama karena komposisi usianya sekitar 56% pemain Roblox berada di bawah usia 17 tahun [4]. Sementara mayoritas server publik tidak menerapkan segregasi usia yang memadai. Anak-anak pun berinteraksi bebas dalam satu ruang percakapan yang sama dengan pengguna dewasa, tanpa filter sosial yang biasanya membatasi topik dan gaya bahasa yang layak diterima audiens di bawah umur. Konsekuensinya nyata, Komisi Perlindungan Anak Indonesia (KPAI) mencatat bahwa interaksi tanpa batas semacam ini kerap berujung pada cyberbullying, ancaman kekerasan verbal, hingga manipulasi psikologis. Kajian Chawki bahkan menyimpulkan bahwa kegagalan sistem filter moderasi bawaan menjadi salah satu faktor utama di balik paparan bahasa yang merusak mental anak-anak di platform ini [5], sejalan dengan tinjauan sistematis terhadap 32 studi yang mengonfirmasi asosiasi antara perundungan daring dengan depresi pada 15–73% kasus dan kecemasan pada 27–84,1% kasus [6].

Secara institusional, platform Roblox tidak membiarkan penggunanya berekspresi tanpa batas. Berdasarkan panduan Community Standards resmi, Roblox

secara eksplisit melarang ujaran kebencian, diskriminasi, dan pelecehan dalam bentuk apa pun, termasuk pada ruang permainan yang mengizinkan penggunaan kata-kata kasar. Oleh karena itu, penyebaran konten rasis dan harassment yang menjadi fokus penelitian ini pada dasarnya merupakan pelanggaran terhadap regulasi internal Roblox. Persoalannya ada pada lemahnya penegakan, bukan ketiadaan aturan. Akar masalahnya ada pada keterbatasan mesin moderasi tersentralisasi milik Roblox. Filter teks bawaan tersebut pada dasarnya bekerja mirip pencocokan kata kaku, sehingga kesulitan menangani variasi bahasa alami non-Inggris seperti bahasa Indonesia. Niat ujaran kebencian yang disamarkan lewat slang, singkatan, atau manipulasi alfanumerik, misalnya "gbl00k", kerap lolos begitu saja tanpa terdeteksi. Di sisi lain, ketidakpahaman konteks lokal ini juga memicu masalah ke arah sebaliknya, yaitu kesalahan sensor, di mana percakapan yang sebenarnya wajar justru berubah menjadi deretan tagar (#####) tanpa alasan logis. Akibatnya, kelancaran komunikasi dan pengalaman bermain pemain jadi terganggu [7], [8]. Celah antara aturan yang sudah ada dengan lemahnya penegakan inilah yang menjadi titik tolak penelitian ini dibutuhkan pendekatan deteksi yang lebih peka konteks daripada sekadar pencocokan kata kunci.

Large Language Model menawarkan jalan keluar dari titik buta tersebut. Albladi et al. menyimpulkan bahwa arsitektur Transformer dan LLM kini menjadi standar dalam pemrosesan bahasa berkat ketangguhannya memahami konteks skala besar [9], Gomis dan Arteta membuktikan performa LLM keluarga Qwen mengungguli model klasik seperti BERT dalam mendeteksi ancaman dan ujaran kebencian [10]. Keunggulan Qwen tersebut juga divalidasi lebih lanjut oleh Pholphuksa *et al.*, yang menemukan bahwa arsitektur Qwen memiliki nilai Recall dalam kepekaan menangkap bahasa kotor yang sangat tajam [11]. Atas dasar itu, penelitian ini memilih Qwen3.5-4B, yang dilatih dengan triliunan token linguistik Asia sehingga relatif adaptif terhadap konteks bahasa Indonesia. Model ini juga tergolong Small Language Model yang, menurut Zhan et al., mampu menyamai performa moderasi model berukuran jauh lebih besar dengan beban komputasi yang jauh lebih ringan, sehingga ideal untuk operasional real-time di server lokal [12].

Untuk benar-benar mencegah konten toksik sampai ke pengguna lain, bukan

sekadar menindak setelah pesan terlanjur tampil, penelitian ini mengusulkan pendekatan pre-moderation proaktif. Antarmuka obrolan bawaan Roblox digantikan dengan antarmuka kustom yang menahan tiap pesan untuk dievaluasi oleh server AI lokal terlebih dahulu, pesan yang aman diteruskan, sedangkan yang toksik disensor sebelum sempat dipublikasikan. Proses ini didukung oleh fase data cleaning dan rekayasa instruksi prompt engineering agar LLM dapat menembus manipulasi ketikan tanpa perlu dilatih ulang setiap kali muncul pola baru [13], [14]. Instruksi ini akan memaksa AI untuk memberikan prediksi pasti berupa satu dari tiga label mutlak antara lain *n*eutral, *r*acist, dan *h*arassment. Model yang telah dimuat ke dalam Ollama ini kemudian dihubungkan ke Roblox melalui REST API dan tunneling Ngrok, sehingga skrip Luau di sisi klien dapat memantau chat dan menjatuhkan sanksi dalam hitungan milidetik.

Sebagai fase penyelesaian yang mengadopsi standar metodologi CRISP-DM, tahap akhir dalam penelitian ini adalah evaluasi kinerja dari *prototype* yang telah dibangun. Sistem akan dihitung nilai Accuracy, Precision dan Recall terhadap dataset chat berbahasa Indonesia, dan kinerja jaringan berupa response time serta latensi server saat memproses banyak log pesan secara bersamaan. Kedua pengujian ini dilakukan langsung di lingkungan simulasi Roblox Studio untuk memastikan sistem tidak asal menghukum pengguna sekaligus tetap responsif dalam kondisi mendekati penggunaan nyata.

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana cara mengimplementasikan model *Large Language Model* Qwen3.5-4B pada *local server* serta mengintegrasikannya ke dalam platform *game* Roblox sebagai sistem moderasi *chat* otomatis?
2. Seberapa tinggi tingkat akurasi model Qwen3.5-4B dalam mendeteksi dan mengklasifikasikan pesan obrolan dengan kategori *neutral, racist,* dan *harassment* ketika diuji menggunakan bahasa Indonesia?

3. Bagaimana kinerja kelancaran komunikasi data dan waktu respons dari prototipe sistem moderasi saat mengeksekusi *chat* secara *real-time* menggunakan lokal *server*?

1.3. Batasan Penelitian

Agar penelitian ini tetap terfokus pada tujuan utama dan dapat diselesaikan dengan sumber daya yang tersedia, penulis menetapkan batasan masalah sebagai berikut:

1. Model AI: Menggunakan *Large Language Model* Qwen3.5-4B versi terkuantisasi yang dijalankan menggunakan *framework* Ollama secara lokal.
2. Lingkungan Implementasi: Integrasi dilakukan antara *Roblox Studio* dan *Local Server* Ollama melalui antarmuka *Native REST API* menggunakan bantuan *tunneling* yaitu Ngrok.
3. Fokus Bahasa: Model dilatih dan diuji khusus untuk mendeteksi Bahasa Indonesia, termasuk variasi bahasa gaul atau *slang*, singkatan, dan manipulasi teks.
4. Penelitian ini difokuskan secara eksklusif pada moderasi data berbasis teks dan tidak mencakup moderasi pada obrolan suara.
5. Klasifikasi Kategori: Deteksi dibatasi pada tiga label utama yaitu neutral, racist, dan harassment.
6. Tahapan Metodologi: Implementasi menggunakan kerangka kerja CRISP-DM dibatasi hanya sampai pada tahap Evaluasi. Penelitian tidak mencakup tahap *Deployment* atau maksudnya tidak melakukan *hosting* sistem ke *cloud server* produksi komersial mengingat tingginya biaya operasional sewa *Cloud GPU*. Sistem diimplementasikan dan dibuktikan sebatas skala *prototype* fungsional.
7. Mekanisme Penanganan: Tindakan otomatis yang dilakukan sistem terbatas pada tindakan di dalam permainan, memberikan peringatan visual atau pengurangan poin kesehatan karakter, dan punishment paling tinggi itu kick dari server bukan pemblokiran akun permanen pada level *database* Roblox.

1.4. Tujuan

Berdasarkan rumusan masalah yang telah disusun, tujuan dari penelitian ini adalah:

1. Mengimplementasikan model LLM Qwen3.5-4B pada *local server* dan mengintegrasikannya ke dalam platform Roblox untuk difungsikan sebagai *prototype* sistem moderasi obrolan secara otomatis.
2. Mengetahui tingkat akurasi model Qwen3.5-4B dalam mendeteksi dan mengklasifikasikan teks obrolan pemain ke dalam ketiga kategori yaitu *neutral, racist,* dan *harassment* menggunakan *dataset* uji berbahasa Indonesia.
3. Mengukur dan mengevaluasi kinerja operasional prototipe sistem, meliputi kecepatan waktu respons dan kelancaran komunikasi data, saat dijalankan secara *real-time* di dalam lingkungan permainan.

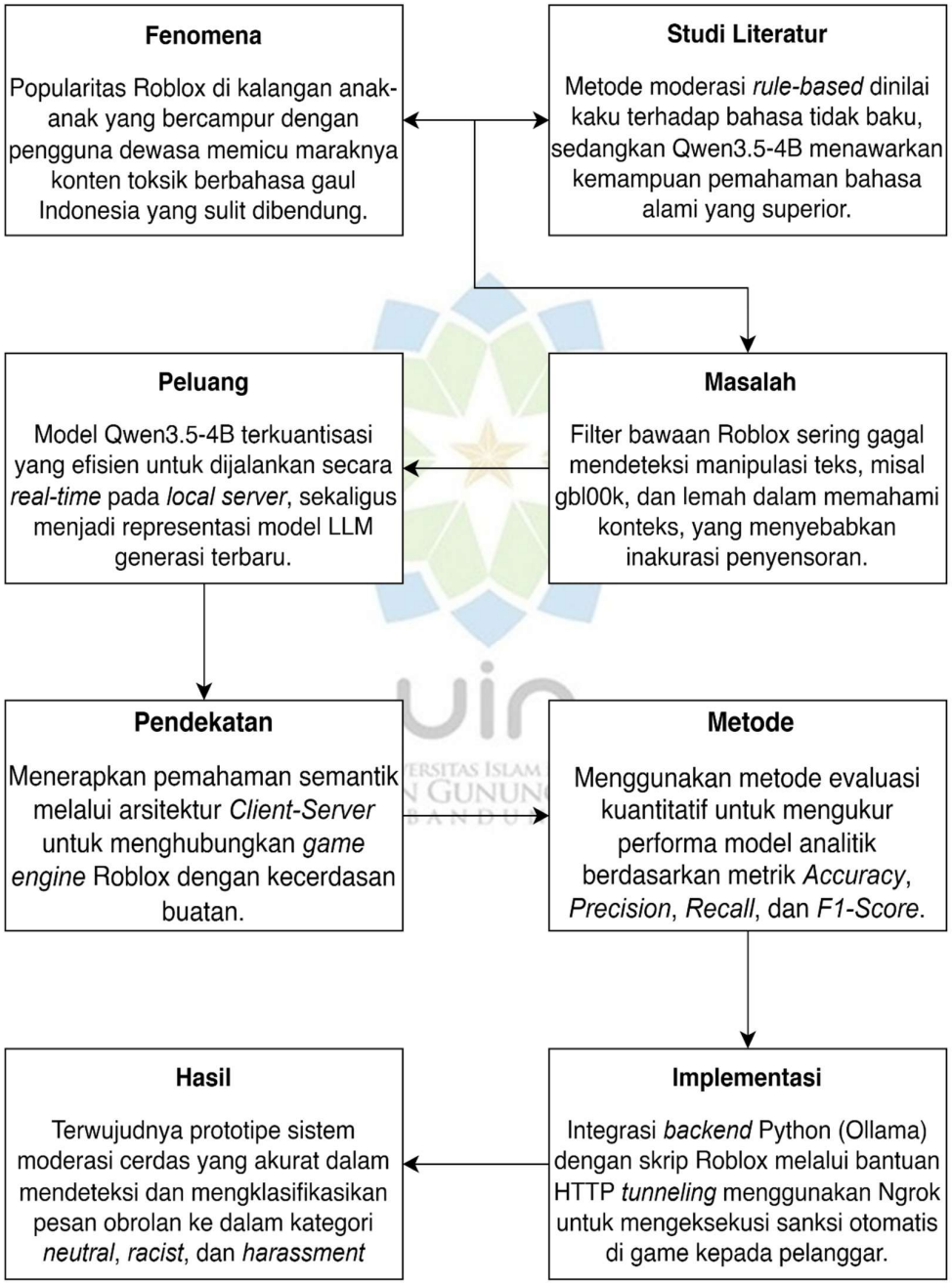
1.5. Manfaat

Manfaat yang bisa didapat dari penelitian ini adalah sebagai berikut :

1. Bagi Pengembang Game: Menyediakan solusi efisiensi sumber daya manusia, di mana pengembang tidak perlu lagi memantau *chat* secara manual 24 jam.
2. Bagi Pemain: Menciptakan lingkungan bermain yang lebih aman dari paparan konten dewasa dan toksik, mengingat tercampurnya demografi usia pemain di Roblox.

1.6. Kerangka Berpikir

Kerangka berpikir yang menggambarkan alur penyelesaian masalah moderasi konten Roblox, mulai dari identifikasi fenomena hingga solusi teknis berbasis LLM Qwen3.5-4B, ditunjukkan secara skematis pada Gambar 1.1



Gambar 1. 1 Kerangka Pemikiran