ABSTRAK

Roblox merupakan platform game dengan basis pengguna yang didominasi anak-anak dan remaja, namun sistem moderasi obrolan bawaannya masih rentan terhadap variasi bahasa gaul, singkatan, dan manipulasi karakter atau *leetspeak* khas bahasa Indonesia, sehingga banyak konten toksik yang lolos dari deteksi. Penelitian ini bertujuan merancang dan mengimplementasikan sistem moderasi obrolan otomatis berbasis *Large Language Model* Qwen3.5-4B pada platform Roblox, mengukur kinerja operasional sistem secara *real-time*, serta mengevaluasi akurasi model dalam mengklasifikasikan teks ke dalam tiga kategori yaitu *neutral*, *racist*, dan *harassment*. Penelitian menggunakan metodologi CRISP-DM yang dibatasi hingga tahap *Evaluation*, meliputi *fine-tuning* model menggunakan metode QLoRA 4-bit pada dataset sebanyak 9.000 baris percakapan berbahasa Indonesia dengan rasio pembagian 80% data latih, 10% data validasi, dan 10% data uji. Model yang telah dilatih diintegrasikan ke platform Roblox melalui layanan *REST API* Ollama dan *HTTP tunneling* Ngrok dengan menerapkan arsitektur *pre-moderation* proaktif menggunakan antarmuka obrolan kustom. Guna menjaga kenyamanan pengalaman bermain, sistem turut dilengkapi mekanisme *Optimistic Local Echo* dan mitigasi *Fail-Open Fallback*. Hasil pengujian menunjukkan model hasil *fine-tuning* mencapai akurasi 87,11%, meningkat drastis dari model dasar yaitu 56,67%. Kinerja operasional sistem mencatatkan latensi rata-rata pemrosesan *API* sebesar 0,58 detik pada kondisi sekuensial dan 0,74 detik pada kondisi *concurrent*. Pengujian fungsional membuktikan sistem berhasil menahan pesan, menyensor konten toksik sebelum dipublikasikan ke publik, serta mengeksekusi sanksi otomatis berupa pengurangan *Health Points* tanpa memicu *lag* pada permainan.

Kata kunci: *Large Language Model, Qwen3.5-4B, moderasi chat, QLoRA, Roblox, klasifikasi teks*

ABSTRACT

Roblox is a gaming platform with a user base predominantly consisting of children and teenagers, yet its built-in chat moderation system remains vulnerable to variations of Indonesian slang, abbreviations, and character manipulation (leetspeak), allowing numerous toxic messages to bypass existing detection mechanisms. This study aims to design and implement an automated chat moderation system based on the Qwen3.5-4B Large Language Model on the Roblox platform, measure operational performance in real-time, and evaluate the model's accuracy in classifying texts into three categories neutral, racist, and harassment. This research employs the CRISP-DM methodology limited to the Evaluation phase, encompassing model fine-tuning using the QLoRA 4-bit method on a dataset of 9.000 rows of Indonesian conversations split into an 80% training set, a 10% validation set, and a 10% test set. The trained model is integrated into the Roblox platform via Ollama REST API services and Ngrok HTTP tunneling, adopting a proactive pre-moderation architecture through a Custom Chat UI. To preserve user experience and interaction comfort, the system is further equipped with Optimistic Local Echo and Fail-Open Fallback mitigation mechanisms. Evaluation results demonstrate that the fine-tuned model achieves an accuracy of 87.11%, improving significantly from the baseline model's 56.67%. Operational performance records an average API processing latency of 0.58 seconds under sequential conditions and 0.74 seconds under concurrent conditions. Functional testing proves that the system successfully intercepts messages, censors toxic content before public publication, and executes automated penalties in the form of Health Points deductions without triggering game lag.

Keywords: *Large Language Model, Qwen3.5-4B, chat moderation, QLoRA, Roblox, text classification*