

ABSTRAK

Sistem penyampaian informasi di SMAN 28 Kabupaten Tangerang masih menghadapi kendala dalam hal kecepatan, ketersediaan, dan konsistensi informasi, terutama pada periode Seleksi Penerimaan Murid Baru (SPMB). Penelitian ini bertujuan untuk merancang dan mengoptimalkan sistem *chatbot* informasi berbasis *website* dengan menerapkan metode *Retrieval-Augmented Generation* menggunakan *framework* LlamaIndex. Metode pengembangan yang digunakan adalah *prototype*, dengan pengujian fungsional menggunakan *Black Box Testing* dan pengujian performa menggunakan Apache JMeter. Optimalisasi kinerja sistem dilakukan melalui penerapan mekanisme *caching query* menggunakan Redis untuk meningkatkan efisiensi pemrosesan dan respons sistem. Hasil pengujian fungsional menunjukkan bahwa seluruh fitur sistem berjalan sesuai spesifikasi dengan tingkat keberhasilan sebesar 100%. Pada pengujian performa, penerapan *caching* mampu menurunkan rata-rata waktu respons dari 18.392 ms menjadi 53 ms pada skenario 30 virtual users, serta meningkatkan *throughput* dari 1,1 *request*/detik menjadi 3,1 *request*/detik. Sistem mampu mempertahankan performa yang layak hingga 700 virtual users, namun mulai mengalami *error rate* sebesar 0,85% pada beban 800 virtual users. Hal ini mengindikasikan adanya keterbatasan sistem dalam menangani *spike load* secara mendadak pada penerapan arsitektur *monolith* dan lingkungan server lokal berbasis Docker. Penelitian ini menunjukkan bahwa penerapan metode *Retrieval-Augmented Generation* menggunakan *framework* LlamaIndex efektif dalam pengembangan *chatbot* informasi sekolah yang responsif, stabil, dan layak digunakan.

Kata Kunci : *Chatbot, RAG, Llama Index, Caching Query.*

ABSTRACT

The information delivery system at SMAN 28 Kabupaten Tangerang still faces challenges in terms of speed, availability, and information consistency, particularly during the New Student Admission Selection (SPMB) period. This study aims to design and optimize a web-based informational chatbot system by implementing the Retrieval-Augmented Generation (RAG) method using the LlamaIndex framework. The development method employed is the prototype approach, with functional testing conducted using Black Box Testing and performance testing using Apache JMeter. System performance optimization is achieved through the implementation of a query caching mechanism using Redis to improve processing efficiency and system response time. Functional testing results indicate that all system features operate according to specifications with a 100% success rate. Performance testing shows that caching reduces the average response time from 18,392 ms to 53 ms in the 30 virtual users scenario and increases throughput from 1.1 requests/second to 3.1 requests/second. The system maintains acceptable performance up to 700 virtual users but begins to experience an error rate of 0.85% at a load of 800 virtual users. This indicates system limitations in handling sudden spike loads under a monolithic architecture and a Docker-based local server environment. The results demonstrate that the implementation of the Retrieval-Augmented Generation method using the LlamaIndex framework is effective for developing a responsive, stable, and reliable school information chatbot.

Keywords: Chatbot, RAG, Llama Index, Caching Query.