ABSTRAK

Hotel Forriz Yogyakarta saat ini masih mengandalkan layanan *WhatsApp Business* dalam menjawab pertanyaan dari tamu sehingga masih ketergantungan oleh manusia. Untuk meningkatkan kualitas pelayanan, dibutuhkan solusi berupa *chatbot* generatif yang mampu memberikan informasi secara otomatis, terkini, dan dinamis. *Chatbot* yang bersifat generatif membutuhkan dukungan dari *Large Language Model* (LLM) agar dapat memahami dan menjawab berbagai pertanyaan dengan fleksibel. Namun, LLM memiliki keterbatasan karena tidak memiliki akses langsung ke sumber informasi spesifik milik hotel, sehingga jawaban yang dihasilkan tidak relevan dan berpotensi halusinasi jawaban.

Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan penerapan teknik *Retrieval Augmented Generation* (RAG) dalam pengembangan *chatbot* berbasis LLM agar dapat memiliki informasi seputar Hotel Forriz Yogyakarta sehingga mengatasi potensi halusinasi. Proses RAG terdiri dari proses *indexing, retrieval,* dan *generation*. Saat pengguna mengajukan pertanyaan, sistem akan mengambil potongan informasi (*chunks*) yang relevan, lalu digunakan oleh LLM untuk menghasilkan jawaban yang sesuai. Sistem kemudian diuji dengan skenario parameter pada bagian *retrieval* dan *generation* untuk mengetahui pengaruhnya terhadap kinerja dari RAG. Evaluasi dilakukan menggunakan *framework* RAGAS dengan 15 *test set* yang berisi input pengguna, jawaban, referensi (*ground truth*), dan konteks hasil *retrieval*.

Hasil penelitian menunjukkan bahwa penerapan teknik RAG berhasil meningkatkan relevansi jawaban serta mengatasi potensi halusinasi jawaban dari LLM. Dari skenario percobaan yang dilakukan, ditemukan bahwa nilai parameter pada *retrieval* dan *generation* memengaruhi sistem RAG. Selain itu, teknik *character-based chunking* maupun *structure-based chunking* pada proses *indexing* dapat menghasilkan *chunks* yang relevan. Kemudian, evaluasi terhadap 15 *test set* menunjukkan hasil rata-rata metrik *Context Precision* sebesar 0.889, *Context Recall* sebesar 1.000, *Faithfulness* sebesar 0.882, *dan Answer Relevancy* sebesar 0.851. Skor tersebut menunjukkan bahwa sistem mampu mengambil konteks yang relevan, menghasilkan jawaban yang sesuai dengan *ground truth*, serta relevan terhadap pertanyaan. Temuan ini juga diperkuat oleh hasil validasi langsung dari *stakeholder* Hotel Forriz, yang menunjukkan bahwa 15 jawaban dari *test set* yang dihasilkan oleh sistem dengan teknik RAG dinilai sesuai dengan *ground truth*.

Kata Kunci: RAG, LLM, *Chatbot,* Halusinasi

ABSTRACT

Hotel Forriz Yogyakarta currently relies on WhatsApp Business to respond to guest inquiries, making it heavily dependent on human intervention. To improve service quality, a solution in the form of a generative chatbot is needed, which can automatically provide up-to-date and dynamic information. A generative chatbot requires support from a Large Language Model (LLM) to flexibly understand and answer various questions. However, LLM have limitations as they do not have direct access to the hotel's specific information sources, which may result in irrelevant or hallucinated answers.

To address this issue, this study proposes the implementation of Retrieval Augmented Generation (RAG) techniques in the development of an LLM-based chatbot so it can access information related to Hotel Forriz Yogyakarta, thus minimizing the potential for hallucinated answers. The RAG process consists of indexing, retrieval, and generation. When a user submits a question, the system retrieves relevant pieces of information, which are then used by the LLM to generate appropriate answers. The system was tested using parameter scenarios in both retrieval and generation stages to evaluate their impact on RAG performance. Evaluation was conducted using the RAGAS framework with 15 test sets consisting of user input, generated responses, references (ground truth), and retrieved context.

The findings indicate that the implementation of RAG enhances the relevance of responses and helps reduce hallucination risks when using LLMs. Experimental scenarios revealed that parameter values in both retrieval and generation stages influence system performance. Furthermore, both character-based and structure-based chunking techniques during indexing produced relevant chunks. Evaluation across the 15 test sets resulted in average scores of 0.889 for Context Precision, 1.000 for Context Recall, 0.882 for Faithfulness, and 0.851 for Answer Relevancy. These scores indicate that the system effectively retrieves relevant context, generates answers consistent with the ground truth, and maintains response relevance. These findings are also supported by direct validation from Hotel Forriz stakeholders, who confirmed that the 15 answers generated by the RAG-based system matched the expected ground truths.

Keywords: RAG, LLM, *Chatbot, Hallucination*