INTISARI

QUESTIONS ANSWERING SYSTEM UNTUK LAYANAN INFORMASI STUNTING BERBASIS OPEN SOURCE LANGUAGE MODEL

Oleh :

Carica Deffa Yullinda

20/462179/PA/20151

Penelitian ini bertujuan untuk membandingkan performa sistem *question answering* berbasis RAG dengan sistem *question answering* berbasis *retrieval* untuk topik *stunting* dengan *open source* LLM. Pada dasarnya, *Large Language Model* mampu memberikan informasi berdasarkan data yang telah dipelajari selama pelatihan sehingga memiliki batasan dalam hal informasi terbaru seperti informasi tentang *stunting*. Maka, sistem ini mengadopsi kemampuan *Retrieval Augmented Generation* mengambil data dari sumber eksternal yaitu *website* Cegah Stunting.

Framework LangChain digunakan untuk mengembangkan 2 sistem RAG yaitu *Retrieval QA Chain* dan *Conversational Retrieval Chain* yang dilengkapi dengan penggunaan memori. Sistem RAG tersebut melibatkan penggunaan LLM *open source* yang meliputi Flan-t5 small, Flan-t5 base, dan Flan-t5 large. Sebagai pembanding, digunakan sistem *Retrieval-Based Question Answering* yang merupakan sistem *retrieval* biasa tanpa LLM.

Hasil pengujian dengan 15 pasangan pertanyaan dan jawaban menunjukkan bahwa *Retrieval Augmented Generation* memiliki performa yang lebih baik dibandingkan sistem *retrieval* biasa. Flan-t5 large mampu menjawab pertanyaan mandiri melalui *Retrieval QA Chain*, tetapi kurang efektif dalam menjawab pertanyaan yang saling berkaitan melalui *Conversational Retrieval Chain*. Flan-t5 large menjadi LLM yang paling optimal dengan perolehan skor metriks yang lebih unggul pada daripada Flan-t5 small dan Flan-t5 base, meskipun waktu responsnya lebih tinggi. *Retrieval QA Chain* memiliki waktu respons yang lebih singkat dibandingkan *Conversational Retrieval Chain*. *Chunk size* dan *chunk overlap* berpengaruh pada waktu dan respons dari sistem *question answering*.

Kata Kunci : Sistem *Question Answering, Large Language Model, LangChain, Retrieval Augmented Generation, Stunting,* BLEU, ROUGE

ABSTRACT

QUESTIONS ANSWERING SYSTEM FOR A STUNTING INFORMATION SERVICE BASED ON OPEN SOURCE LANGUAGE MODELS

By :

Carica Deffa Yullinda

20/462179/PA/20151

This research aims to compare the performance of a RAG-based question answering system with a retrieval-based question answering system for stunting topics with open source LLM. Basically, the Large Language Model is capable of providing information based on data learned during training so it has limitations in terms of up-to-date information such as information about stunting. So, the system adopts the ability of Retrieval Augmented Generation to retrieve data from an external source called Cegah Stunting website.

The LangChain framework is used to develop two RAG systems, the Retrieval QA Chain and the Conversational Retrieval Chain, which are equipped with memory usage. The RAG system involves the use of open source LLMs which include Flan-t5 small, Flan-t5 base, and Flan-t5 large. As a comparison, the Retrieval-Based Question Answering system is used, which is an ordinary retrieval system without LLM.

Test results with 15 pairs of questions and answers show that Retrieval Augmented Generation has better performance than ordinary retrieval systems. Flan-t5 large is able to answer independent questions via Retrieval QA Chain, but is less effective in answering interrelated questions via Conversational Retrieval Chain. Flan-t5 large is the most optimal LLM with superior metric scores compared to Flan-t5 small and Flan-t5 base, even though the response time is higher. Retrieval QA Chain has a shorter response time than Conversational Retrieval Chain. Chunk size and chunk overlap affect the time and response of the question answering system.

Keywords: Question Answering System, Large Language Model, LangChain, Retrieval Augmented Generation, Stunting, BLEU, ROUGE