

Development of a Metadata-Related Benchmark Based on Large Language Models (LLMs) for the Retrieval-Augmented Generation (RAG) Module of Pneuma: A Data Discovery System = Pengembangan Tolok Ukur Terkait Metadata Berbasis Large Language Model (LLM) untuk Modul Retrieval-Augmented Generation (RAG) pada Pneuma: Sistem Pencarian Data

Muhammad Imam Luthfi Balaka, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920568601&lokasi=lokal>

Abstrak

Pneuma is a data discovery system developed to address the limitations of current dataset retrieval methods. It leverages Retrieval-Augmented Generation (RAG) to store and retrieve metadata and dataset summaries. Metadata provides basic dataset information, while summaries include column names and row values. To optimize RAG components, an appropriate benchmark is needed. This work introduces a metadata-based benchmark consisting of triples (T, Q, A), where T is a dataset, Q is a metadata-related question, and A is the corresponding answer based on T. The datasets and questions come from existing sources, while answers are generated and evaluated using two Large Language Models (LLMs): LLM Generator and LLM Evaluator. Experiments identify the best model and strategy, yielding the optimal combination of OpenHermes-2.5-Mistral-7B with a role-play prompt and greedy search decoding. This combination is used to tune Pneuma's RAG module.

.....Pneuma adalah sistem pencarian data yang dikembangkan untuk mengatasi keterbatasan sistem pencarian dataset saat ini. Sistem ini menggunakan mekanisme Retrieval-Augmented Generation (RAG) untuk menyimpan dan mengambil informasi, termasuk metadata dan ringkasan dataset. Metadata menjelaskan informasi dasar dataset, sedangkan ringkasan dataset mencakup nama kolom dan nilai baris. Untuk memilih dan menyetel komponen RAG, diperlukan tolok ukur yang sesuai. Skripsi ini mengembangkan tolok ukur berbasis metadata dalam bentuk tripel (T, Q, A), dengan T sebagai dataset, Q sebagai pertanyaan metadata, dan A sebagai jawaban berdasarkan T. Dataset dan pertanyaan diperoleh dari sumber yang ada, sementara jawaban dihasilkan dan dievaluasi menggunakan dua Large Language Models (LLM), yaitu LLM Generator dan LLM Evaluator. Eksperimen dilakukan untuk memilih model dan strategi terbaik, menghasilkan kombinasi optimal OpenHermes-2.5-Mistral-7B dengan prompt role-play, serta strategi dekoding greedy search. Kombinasi ini digunakan untuk menyetel modul RAG pada Pneuma.