UNIVERSITAS
GADJAH MADA
ANALISIS PENGARUH PARAMETER RETRIEVAL-AUGMENTED GENERATION PADA LARGE LANGUAGE MODEL UNTUK SISTEM DIAGNOSIS BANDING CEDERA OLAHRAGA LARI
Muhamad Bintang Harry Dharmawan, Dr. Eng. Silmi Fauziati, S.T., M.T.; Ir. Azkario Rizky Pratama, S.T., M.Eng., Ph
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Olahraga lari telah menjadi aktivitas fisik yang paling banyak diminati di Indonesia, tetapi peningkatan minat yang pesat juga membawa lonjakan kasus cedera muskuloskeletal terkait lari (*Running-related Musculoskeletal Injuries*). Disisi lain, diagnosis banding cedera olahraga lari merupakan proses yang menantang karena memiliki gejala yang tumpang tindih, sementara dokter dihadapkan pada kasus yang banyak dan beban kognitif yang tinggi karena meningkatnya insiden cedera lari. Meskipun *Large Language Model* (LLM) menawarkan potensi sebagai komponen pada Sistem Pendukung Keputusan Klinis, implementasinya pada domain medis spesifik terhambat oleh kecenderungan berhalusinasi akibat kurangnya pelatihan parametrik. Penelitian ini mengusulkan integrasi arsitektur *Retrieval-Augmented Generation* (RAG) dan *few-shot prompting* pada model Qwen2.5:3b-instruct untuk menghasilkan diagnosis banding cedera olahraga lari.

Eksperimen sistematis dilakukan terhadap 18 konfigurasi parameter *pipeline* RAG yang mencakup variasi *chunk size* (256, 384, 512), *overlap* (0 dan 25% dari *chunk size*), dan top-k (3, 5, 10) menggunakan dataset sintetik 72 *clinical vignette* yang mencakup delapan diagnosis target RRMI. Performa setiap konfigurasi dievaluasi menggunakan *framework* evaluasi empat lapis, meliputi *Context Precision* dari *framework* RAGAS, metrik berbasis ranking (Hit@1, Hit@3, *Mean Reciprocal Rank*), distribusi gradasi kemiripan semantik (*Semantic Answer Similarity*), serta evaluasi klinis oleh tiga dokter terhadap lima dimensi penilaian: *Appropriateness, Comprehensiveness, Clinical Reasoning Quality, Patient Safety,* dan *Clinical Usefulness*.

Hasil eksperimen menunjukkan bahwa rata-rata 18 konfigurasi RAG mengungguli kondisi *baseline* pada seluruh metrik *ranking-based*. Konfigurasi terbaik yang teridentifikasi adalah *chunk size* = 256, *overlap* = 0, dan top-k = 10, dengan Hit@1 = 72,22%, Hit@3 = 84,72%, MRR = 0,7755. *Chunk size* terbukti sebagai parameter paling berpengaruh, sementara *overlap* memberikan pengaruh yang minimal. Temuan penting bahwa *Context Precision* berkorelasi negatif dengan performa metrik *ranking-based*, menegaskan bahwa evaluasi kualitas *retrieval* saja tidak memadai sebagai proksi kualitas output akhir sistem. Evaluasi klinis pada konfigurasi terbaik mengungkapkan bahwa dimensi *Patient Safety* mendapatkan skor tertinggi dan *Clinical Reasoning Quality* mendapatkan skor terendah. Analisis korelasi *Spearman* antara metrik otomatis dan evaluasi klinis menunjukkan korelasi yang lemah hingga sedang, mengonfirmasi bahwa evaluasi klinis oleh dokter tidak dapat digantikan oleh metrik otomatis.

Kata kunci : *retrieval-augmented generation, large language model,* diagnosis banding, cedera olahraga lari, sistem pendukung keputusan klinis

UNIVERSITAS
GADJAH MADA
ANALISIS PENGARUH PARAMETER RETRIEVAL-AUGMENTED GENERATION PADA LARGE LANGUAGE MODEL UNTUK SISTEM DIAGNOSIS BANDING CEDERA OLAHRAGA LARI
Muhamad Bintang Harry Dharmawan, Dr. Eng. Silmi Fauziati, S.T., M.T.; Ir. Azkario Rizky Pratama, S.T., M.Eng., Ph
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Running has become the most widely pursued physical activity in Indonesia; however, the rapid growth in participation has also brought a significant rise in running-related musculoskeletal injuries (RRMI). At the same time, differential diagnosis of running injuries is a particularly challenging process due to overlapping symptoms across conditions, while clinicians face high caseloads and increasing cognitive demands driven by the rising incidence of running injuries. Although Large Language Models (LLMs) offer promising potential as components of Clinical Decision Support Systems, their application to specific medical domains is hampered by a tendency toward hallucination resulting from insufficient parametric training. This study proposes the integration of Retrieval-Augmented Generation (RAG) architecture and few-shot prompting with the Qwen2.5:3b-instruct model to generate differential diagnoses for running-related musculoskeletal injuries.

A systematic experiment was conducted across 18 RAG pipeline parameter configurations, varying chunk size (256, 384, 512), overlap (0 and 25% of chunk size), and top-k (3, 5, 10) using a synthetic dataset of 72 clinical vignettes covering eight target RRMI diagnoses. Each configuration was evaluated using a four-layer evaluation framework comprising Context Precision from the RAGAS framework, ranking-based metrics (Hit@1, Hit@3, Mean Reciprocal Rank), Semantic Answer Similarity gradient distribution, and clinical evaluation by three physicians across five assessment dimensions: Appropriateness, Comprehensiveness, Clinical Reasoning Quality, Patient Safety, and Clinical Usefulness.

Results demonstrate that the average of 18 RAG configurations outperformed the baseline condition across all ranking-based metrics. The optimal configuration identified was chunk size = 256, overlap = 0, and top-k = 10, achieving Hit@1 = 72.22%, Hit@3 = 84.72%, and MRR = 0.7755. Chunk size proved to be the most influential parameter, while overlap had minimal impact. A key finding is that Context Precision correlated negatively with ranking-based metric performance, demonstrating that retrieval quality alone is insufficient as a proxy for end-to-end system output quality. Clinical evaluation of the best configuration revealed that Patient Safety received the highest score and Clinical Reasoning Quality received the lowest. Spearman correlation analysis between automatic metrics and clinical evaluation scores showed weak-to-moderate correlations, confirming that physician clinical evaluation cannot be replaced by automatic metrics alone.

Keywords : *retrieval-augmented generation, large language model, diagnosis banding, running-related injuries, clinical decision support system*