

## INTISARI

Model Bahasa Besar (LLM) menunjukkan potensi signifikan dalam diagnosis dan penalaran klinis, namun keandalan model tanpa penyesuaian khusus seringkali masih terbatas oleh kurangnya pengetahuan domain spesifik. Penelitian ini secara sistematis mengevaluasi teknik *prompt engineering*, *fine-tuning*, dan *retrieval-augmented generation* (RAG) untuk mengoptimalkan kinerja diagnostik sistem berbasis LLM. Menggunakan Llama 3.2 3B Instruct sebagai model dasar, penelitian ini pertama-tama mengidentifikasi prompt *chain-of-thought* yang paling efektif sebagai *baseline*. Kinerja model dasar tersebut kemudian ditingkatkan dengan menggunakan RAG yang memanfaatkan sumber pengetahuan klinis eksternal serta *parameter-efficient fine-tuning* menggunakan QLoRA pada data medis domain-spesifik. Enam konfigurasi optimasi dibandingkan secara komprehensif.

Kinerja model dievaluasi dengan menggunakan metrik kuantitatif dan kualitatif. Metrik kuantitatif otomatis meliputi Perplexity, BERTScore, dan Semantic Answer Similarity. Sementara itu, evaluasi kualitatif dilakukan melalui tinjauan oleh lima tenaga medis berdasarkan kerangka QUEST, yang menilai aspek Akurasi, Penalaran, Kejelasan, Keamanan, dan Kepercayaan. Data hasil tinjauan kualitatif ini kemudian divalidasi dengan menggunakan analisis statistik.

Hasil penelitian menunjukkan bahwa strategi terintegrasi yang menggabungkan *prompt engineering*, *fine-tuning*, dan RAG secara konsisten memberikan kinerja terbaik. Pendekatan ini mencapai skor Perplexity terendah dan Semantic Answer Similarity tertinggi dalam evaluasi otomatis, sementara metode berbasis *fine-tuning* menunjukkan keunggulan pada metrik BERTScore. Dalam evaluasi oleh tenaga medis, para ahli menilai konfigurasi terintegrasi penuh ini secara signifikan lebih unggul pada aspek akurasi dan keamanan dibandingkan konfigurasi lainnya, meskipun terdapat catatan bahwa model cenderung menghasilkan respons yang terlalu panjang atau mengandung pengulangan. Temuan ini mengukuhkan bahwa integrasi sinergis antara *prompt engineering* yang dioptimalkan, *fine-tuning* pada data domain-spesifik, dan augmentasi RAG merupakan solusi paling efektif untuk meningkatkan akurasi dan persepsi keamanan dalam sistem diagnosis klinis berbasis LLM.

**Kata Kunci:** Diagnosis klinis, model bahasa besar, *prompt engineering*, *fine-tuning*, *retrieval-augmented generation*

## ABSTRACT

Large Language Models (LLMs) show significant potential for clinical diagnosis and reasoning, yet standard models often face reliability issues. This research systematically evaluates prompt engineering, fine-tuning, and Retrieval-Augmented Generation (RAG) to optimize LLM diagnostic capabilities. The Llama 3.2 3B Instruct model served as the baseline, with an effective initial prompt established using Intuitive Reasoning Chain-of-Thought. This research then enhanced this baseline with retrieval-augmented generation using external clinical knowledge and with parameter-efficient fine-tuning on domain-specific medical data. This research assessed six configurations—individual techniques, pairwise combinations, and the fully integrated approach.

Performance was evaluated using two complementary methods. Automated metrics included Perplexity, BERTScore, and Semantic Answer Similarity. Concurrently, five medical professionals conducted qualitative expert reviews based on the QUEST framework, assessing Accuracy, Reasoning, Clarity, Safety, and Trust. Statistical analyses were applied to validate the human evaluation data.

Results from automated metrics indicated that combined prompt engineering, fine-tuning, and RAG approach consistently achieved the best balance, yielding the lowest Perplexity and highest Semantic Answer Similarity. Fine-tuning based methods excelled in BERTScore. Human evaluations corroborated these findings, showing that the fully integrated strategy was rated significantly higher in accuracy and safety. Although qualitative feedback was positive, reviewers noted occasional verbosity even in advanced configurations. This study concludes that a synergistic integration of optimized prompt engineering, domain-specific fine-tuning, and retrieval-augmented generation provides the most robust solution for enhancing LLM-based clinical diagnosis, markedly improving accuracy and perceived safety.

**Keywords:** Clinical diagnosis, large language model, prompt engineering, fine-tuning, retrieval-augmented generation