

Multi-Lingual Malaysian Embedding: Leveraging Large Language Models for Enhanced Semantic Representations

Husein Zolkepli* Aisyah Razak† Kamarul Adha‡ Ariff Nazhan§

January 22, 2024

Abstract

In this work, we present a novel approach to multi-lingual embedding in the context of the Malaysian language. We fine-tuned the Malaysian Llama2 model specifically for embedding tasks involving both negative and positive pairs. The resulting embeddings exhibit exceptional performance, showcasing their suitability for applications related to semantic similarity and RAG (Retrieval-Augmented Generation).

Our approach demonstrates state-of-the-art results, surpassing the performance of OpenAI's text-embedding-ada-002 model on Malaysian contextual embedding and RAG tasks. The fine-tuned Malaysian Llama2 model not only sets a new benchmark but also establishes itself as a powerful tool for capturing nuanced semantic relationships in the multi-lingual Malaysian language landscape.

All models released at [HuggingFace Mesolitica Malaysian Embedding Collection](#).

*husein@mesolitica.com

†aisyahrazak171@gmail.com

‡kamarul.adha360@gmail.com

§ariffnzhn@gmail.com