MaLLaM Malaysia Large Language Model

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

January 20, 2024

Abstract

Addressing the gap in Large Language Model pretrained from scratch with Malaysian context, We trained models with 1.1 billion, 3 billion, and 5 billion parameters on a substantial 349GB dataset, equivalent to 90 billion tokens based on our pretrained Byte Pair Encoding (BPE) tokenizer for a single epoch. MaLLaM contributes to enhanced natural language understanding and generation tasks in the Malay language.

MaLLaM models mark a significant contribution to the field, providing comprehensive language representations grounded in Malaysian context. This endeavor aims to pave the way for enhanced natural language understanding and generation tasks specific to the linguistic nuances present in Malaysia. We discuss the training methodology, dataset composition, and the potential impact of MaLLaM in advancing the capabilities of large language models within the context of the Malay language.

All models released at HuggingFace Mesolitica MaLLaM Collection.

*husein@mesolitica.com
†aisyahrazak171@gmail.com
‡kamarul.adha360@gmail.com
§ariffnzhn@gmail.com