# The Criteria for Large Language Model (LLM) Testing Framework

**Zulkefli Mansor**
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

**Rodziah Latih**
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

**Xiaoyan Zhao**
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

**Kamalrufadillah Sutling**
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

**Badrisham Ahmad**
Center for Information Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

**Ahmad Najmi Ismail**
Center for Information Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

## Abstract

The rapid proliferation of Large Language Models (LLMs) across education, business, and research domains has underscored the urgent need for systematic evaluation methods that go beyond traditional accuracy benchmarks. This paper presents a comprehensive LLM Testing Framework designed to assess large-scale language models across eight multidimensional parameters: core model performance, linguistic and communication quality, reasoning and comprehension, safety, ethics and bias, system and functional performance, user and task satisfaction, domain specificity, and maintainability and monitoring. The framework integrates both quantitative and qualitative metrics including factual accuracy, coherence, bias detection, privacy safeguards, latency, and human-centered trust evaluations to provide a holistic view of model capabilities and risks. A five-point scoring scale enables consistent benchmarking and facilitates performance comparisons between models or configurations. Unlike existing evaluation tools that emphasize task completion or benchmark accuracy alone, this framework emphasizes contextual reliability, ethical compliance, and end-user experience as integral components of LLM quality. Preliminary testing demonstrates that this structured approach supports more transparent, reproducible, and adaptable assessment of LLMs across diverse domains. The proposed framework contributes a standardized yet flexible methodology for researchers, developers, and policymakers seeking to ensure that language models are not only powerful, but also safe, fair, and contextually aligned.

## Keywords

Large Language Models, Evaluation Framework, Testing, Safety, Ethics, Benchmarking, Malay LLM, AI-Warisan.