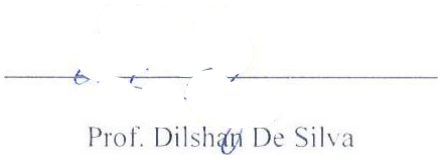
A Metric-Driven Framework for Evaluating Large Language Models in Software Testing: Insights from Industry Experts

V. I. T. Perera
(MS24016520)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION SYSTEMS

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Prof. Dilshan De Silva

Approved for MSc. Research Project:

MSc in IS Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign:  .....

Perera V. I. T.

Date: 27/12/2025

ABSTRACT

A Metric-Driven Framework for Evaluating Large Language Models in Software Testing: Insights from Industry Experts

Isuru Perera
MSc. in Information Systems
Supervisor: Prof. Dilshan De Silva
December 2025

The integration of Large Language Models (LLMs) into software testing workflows has introduced new opportunities for automation, but also raised critical questions regarding the reliability, maintainability, and effectiveness of the generated test cases. This study addresses the lack of standardized evaluation practices by proposing and validating a comprehensive metric-driven framework: STEAM-LLM (Software Test Effectiveness Assessment Model for LLM-generated tests). Through a mixed-methods research design involving structured surveys and expert interviews, the framework identifies three core independent variables: Prompt Engineering Level, Human Intervention, and Input Specification Quality and evaluates their impact on Fault-Revealing Power and Maintainability & Stability. The inclusion of Edit Effectiveness as a mediator, along with contextual moderators such as Task Complexity, Developer Experience, and LLM Class, reflects the nuanced dynamics of LLM-assisted testing. Confounding variables including Baseline Project/Test Quality and Time Spent on Understanding the Task were also accounted for to ensure valid assessments. The framework was empirically validated and supported by metric thresholds (e.g., ≥80% coverage, ≤10% smell density), offering practical benchmarks for industry adoption. The findings demonstrate that structured prompts, strategic human oversight, and high-quality inputs are essential for producing reliable and maintainable LLM-generated test cases. This research contributes a theoretically robust and practically applicable model for evaluating AI-assisted software testing, laying the foundation for future experimentation, tooling, and integration into continuous development environments.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Prof. Dilshan De Silva, for his continuous support, insightful guidance, and constructive feedback throughout the course of this research at Sri Lanka Institute Information Technology. I am also deeply thankful to all the survey and interview participants whose valuable experiences and perspectives greatly enriched the development and validation of the evaluation framework. Finally, I extend my appreciation to everyone who informally supported me during this journey, offering encouragement, ideas, and feedback that helped shape this work.

TABLE OF CONTENTS

List of Figures

List of Tables