CONTENTS

HALAMAN PENGESAHAN i
STATEMENT iii
PAGE OF DEDICATION iv
PREFACE v
CONTENTS vi
LIST OF TABLES ix
LIST OF FIGURES x
NOMENCLATURE AND ABBREVIATION xii
INTISARI xiii
ABSTRACT xiv
CHAPTER I Introduction 1
1.1 Research Background 1
1.2 Problem Statements 3
1.3 Research Objectives 3
1.4 Scope and Limitations 3
1.5 Research Benefits 5
1.6 Structure of Thesis 5
CHAPTER II Literature Review 7
2.1 Prompt Engineering for Clinical Reasoning with Large Language Models 7
2.2 Fine-Tuning for Large Language Model Optimization 9
2.3 Retrieval-Augmented Generation for Large Language Model Optimization 11
2.4 Human Evaluation Framework for Large Language Model in Healthcare .. 14
2.5 Comparative Analysis of the Previous Works 17
CHAPTER III Theoretical Framework 21
3.1 Artificial Intelligence 21
3.2 Machine Learning 21
3.3 Deep Learning 22
3.4 Natural Language Processing 22
3.5 Large Language Models 24
3.6 LLaMA Architecture 24
3.7 Prompt Engineering 27
3.7.1 Zero-shot Prompting 27
3.7.2 Few-shot Prompting 28
3.7.3 Chain-of-Thought (CoT) Prompting 28
3.8 Fine-Tuning 29
3.8.1 Parameter-Efficient Fine-Tuning 30

3.8.2 Low-rank Adaptaion (LoRA) 31
3.8.3 Quantized Low-rank Adaptation (QLoRA) 32
3.9 Retrieval-Augmented Generation 33
3.9.1 Naive RAG 34
3.9.2 Advanced RAG 35
3.9.3 Modular RAG 35
3.10 Perplexity 35
3.11 BERTScore 36
3.12 Semantic Answer Similarity 37
CHAPTER IV Research Methodology 39
4.1 Hardware and Software Specifications 39
4.2 Materials and Data Sources 40
4.3 Research Workflows 40
4.3.1 Literature Review 40
4.3.2 Model Selection 41
4.3.3 Dataset Selection 42
4.3.3.1 Data Preprocessing 43
4.3.4 Prompt Engineering Method 44
4.3.5 Fine-Tuning Method 45
4.3.6 Retrieval-Augmented Generation Method 50
4.3.6.1 Knowledge Base Documents 51
4.3.6.2 Retrieval-Augmented Generation Process 52
4.3.7 Evaluation Method 54
4.3.7.1 Automated Evaluation Metrics 54
4.3.7.2 Human Evaluation 55
4.3.7.3 Statistical Analysis 56
CHAPTER V Result and Discussion 58
5.1 Automated Evaluation Results 58
5.1.1 Baseline Prompt Selection 58
5.1.2 Comparative Automated Evaluation of LLM Optimization 59
5.1.2.1 Perplexity Score Analysis Across Optimized Methods 60
5.1.2.2 BERTScore Analysis Across Optimized Methods 61
5.1.2.3 SAS Score Analysis Across Optimized Methods 62
5.2 Human Evaluation Results 63
5.2.1 Descriptive Statistical 64
5.2.2 Normality Test 65
5.2.3 Inter-Rater Reliability Test 68
5.2.4 Omnibus Test 69
5.2.5 Post-hoc Pairwise Test 69

5.2.6 Qualitative Assessment 71
5.3 Discussion and Summary of Findings 73
5.3.1 Automated Evaluation Summary 73
5.3.2 Human Evaluation Summary 74
CHAPTER VI Conclusion and Future Works 75
6.1 Conclusion 75
6.2 Future Works 75
REFERENCES 77
LAMPIRAN L-1
L.1 Ethical Exemption L-1
L.2 Informed Consent Form L-2
L.3 Questionnaire Form L-7
L.4 Questionnaire Result L-60
L.4.1 Quantitative Result L-60
L.4.2 Quantitative Result L-75
L.5 Preprocessing Data L-89
L.5.1 Example of Nonmedical Domain Questions L-89
L.5.2 Examples of Multiple Choices Questions L-90
L.6 Prompt Engineering Template L-92
L.6.1 Traditional CoT Prompt L-92
L.6.2 Differential Diagnosis CoT Prompt L-93
L.6.3 Intuitive Reasoning CoT Prompt L-94
L.6.4 Analytic Reasoning CoT Prompt L-95
L.6.5 Bayesian Reasoning CoT Prompt L-96
L.7 Pseudocode L-97
L.7.1 Data Preprocessing L-97
L.7.2 Prompt Engineering L-98
L.7.3 Fine-Tuning L-99
L.7.4 Retrieval-Augmented Generation L-100