UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

TABLE OF CONTENTS

TITLE PAGE ..... **ii**
APPROVAL PAGE ..... **iii**
STATEMENT ..... **iv**
ABSTRACT ..... **v**
INTISARI ..... **vi**
ACKNOWLEDGEMENTS ..... **vii**
TABLE OF CONTENTS ..... **viii**
LIST OF TABLES ..... **x**
LIST OF FIGURES AND EQUATIONS ..... **xi**
CHAPTER I INTRODUCTION **1**
1.1 Research Background 1
1.2 Problem Formulation 3
1.3 Research Objective 3
1.4 Research Scope 4
1.5 Research Significance 4
CHAPTER II LITERATURE REVIEW **5**
CHAPTER III RESEARCH METHODOLOGY **12**
3.1. General Description of the Research 12
3.2. Dataset Description 14
3.3. Code Rewriting 15
3.4. Embedding Process 17
3.5. Similarity Measurement 19
3.5.1. Cosine Similarity 20
3.6. Similarity Interpretation and Classification 21
3.7. Evaluation 21
3.8. Experiment Scenario 25
CHAPTER IV EXPERIMENT AND IMPLEMENTATION **26**
4.1. Hardware and Software Environment Usage 26
4.2. Dataset Preparation 27
4.2.1. Initial Dataset Inspection and Column Filtering 28
4.2.2. Data Cleaning and Sampling 28
4.3. SimCSE Training Module 30
4.3.1. Training data for the SimCSE 31
4.3.2. Model Architecture and MLP Projection Head 32
4.3.3. Contrastive Loss 33
4.4. Rewriting Module 34
4.4.1. Model Selection 34
4.4.2. Prompt Design 35
4.4.3. Generation Configuration 35

UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

4.4.4. Batched Generation and Retry Logic 36
4.5. Evaluation Module 38
4.5.1. Batch Sizing and Validity Filtering 38
4.5.2. Similarity Scoring and Metrics 39
4.5.3. Output and Result 40
CHAPTER V RESULT AND DISCUSSION 41
5.1. Result 41
5.1.1. Effect of Rewrite Amount 42
5.1.2. Per-Variant Performance 42
5.2. Discussion 44
5.2.1. CodeT5 vs GraphCodeBERT 45
5.2.2. SimCSE Fine-Tuning Impact 46
5.2.3. Rewrite Amount Impact 46
5.2.4. Variant Instability 47
5.2.5. Comparison with Prior Work 48
5.2.6. Embedding Model Suitability and Selection 48
CHAPTER VI CONCLUSION AND RECOMMENDATION 50
6.1. CONCLUSION 50
6.2. LIMITATIONS 52
6.3. RECOMMENDATION 52
REFERENCES 54

UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

LIST OF TABLES

Table 2.1 Research Comparison Table 10
Table 3.1 Experiment Scenario Table..... 25
Table 4.1 Hardware Specifications..... 26
Table 5.1 Overall Detection Performance AUROC and Pairwise Accuracy (PA) for All Model Configurations..... 41
Table 5.2 Effect of Rewrite Amount (m) on AUROC and Pairwise Accuracy (PA)...42
Table 5.3 Per-Variant AUROC at m = 8 for All Model Configurations.....42
Table 5.4 GraphCodeBERT-BASE m=2 Variant 13 Example.....43
Table 5.5 CodeT5-BASE m=2 Variant 7 Example.....43

UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

LIST OF FIGURES AND EQUATIONS

Figure 3.1. Conceptual Framework 13
Figure 3.2. Variant Type 2 Dataset Code Snippet Example 15
Figure 3.3. Rewriting Prompt 16
Equation 3.1. Embedding Formula 18
Equation 3.2. Similarity Formula 20
Equation 3.3. Cosine Similarity Formula 20
Equation 3.4. Similarity Interpretation and Classification Function 21
Equation 3.5. Resulting Accuracy Formula 22
Equation 3.6. Pairwise Accuracy Formula 23
Equation 3.7. Accuracy Formula 23
Equation 3.8. Precision Formula 23
Equation 3.9. Recall Formula 23
Equation 3.10. F1-Score Formula 24
Equation 3.11. Standard AUROC Definition 24
Equation 3.12. Wilcoxon–Mann–Whitney AUROC Formulation 24
Figure 4.1. Dataset Loading 29
Figure 4.2. Implementation of SimCSE Unsupervised Training 31
Figure 4.3. MLP Projection Head 32
Figure 4.4. SimCSE Loss Function 33
Figure 4.5. Prompt Design 35
Figure 4.6. Rewrite Validity Function 36
Figure 4.7. Padding and Skipping Process 37
Figure 4.8. Score Detection 39