UNIVERSITAS GADJAH MADA
Preserving Functional Fidelity in Compressed Large Language Models Through Corrective and Adaptive Decomposition
Muchammad Daniyal Kautsar, Syukron Abu Ishaq Alfarozi, S.T., Ph.D.;Widyawan, S.T., M.Sc., Ph.D
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

CONTENTS

ENDORSEMENT PAGE i
DECLARATION OF ORIGINALITY 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 Statement 2
1.3 Research Objectives 2
1.4 Scope and Limitations 3
1.5 Research Benefits 5
1.6 Structure of Thesis 5
CHAPTER II Literature Review and Theoretical Framework 7
2.1 Literature Review 7
2.1.1 Fundamentals of Model Compression in LLMs 7
2.1.2 Low-Rank Factorization Techniques in Neural Networks 8
2.1.3 Structured Pruning Techniques 8
2.1.4 Low-Rank Methods for LLM Layer Compression 10
2.1.5 Residual-Based and Adaptive Correction Mechanisms in Model Compression 10
2.1.6 Layerwise and Componentwise Heterogeneity in LLMs 11
2.1.7 Summary of Literature Review 11
2.2 Theoretical Framework 12
2.2.1 Model Efficiency in Modern Artificial Intelligence 12
2.2.2 Deep Learning Fundamentals 13
2.2.3 Large Language Models and Transformer-Based Architectures 13
2.2.4 Mathematical Foundations of Low-Rank Decomposition 15
2.2.4.1 Singular Value Decomposition (SVD) 15
2.2.4.2 Rank Truncation and Low-Rank Approximation 15
2.2.4.3 Matrix Norms and Approximation Quality 16
2.2.4.4 Functional vs. Reconstruction Fidelity 16

UNIVERSITAS
GADJAH MADA
Preserving Functional Fidelity in Compressed Large Language Models Through Corrective and Adaptive Decomposition
Muchammad Daniyal Kautsar, Syukron Abu Ishaq Alfarozi, S.T., Ph.D.;Widyawan, S.T., M.Sc., Ph.D
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

2.3 Comparative Method Analysis 16
2.3.1 Standard SVD-based Compression 17
2.3.2 Structured Pruning 17
2.3.2.1 Width Pruning 17
2.3.2.2 Depth Pruning (Layer Removal) 17
2.3.3 Enhanced SVD and Residual-Based Approaches 18
2.3.4 Justification for the Proposed Corrective Framework 18
2.4 Research Questions 20
CHAPTER III Research Methodology 21
3.1 Research Stages 21
3.2 The CALR Framework 22
3.2.1 High-Level Overview of the CALR Framework 22
3.2.2 Architectural Transformation Details 23
3.2.2.1 Primary SVD-Based Compression of Constituent FFN Layers 24
3.2.2.2 The Additive Low-Rank Corrective Module 24
3.2.3 Formalizing the Corrective Objective 25
3.2.4 Hypothesis on Corrective Module and Empirical Justification 26
3.3 CALR Training, Optimization, and Evaluation Design 27
3.3.1 CALR Training and Optimization Procedure 27
3.3.2 Selecting Candidate Layers for Transformation and Preliminary Analysis on Layer Hetegeronity 28
3.3.2.1 Selective Initialization and Architectural Transformation 31
3.3.2.2 Joint Fine-tuning 32
3.3.3 Experimental Design and Methodology 32
3.3.3.1 Performance Benchmarking 33
3.3.3.2 CALR Configuration and Calibration 34
3.3.3.3 Supervised Fine-Tuning Protocol 34
3.3.3.4 Framework for Statistical Significance Testing 36
3.3.3.5 Framework for Ablation Studies 36
3.3.3.6 Framework for Inference and Latency Analysis 37
CHAPTER IV Empirical Evaluation and Analysis 39
4.1 Performance Comparison with Baselines 39
4.1.1 Analysis on SmolLM2-135M 39
4.1.2 Analysis on Qwen3-0.6B 39
4.1.3 Analysis on Llama-3.2-1B 40
4.2 Statistical Significance of Results 41
4.2.1 Overall Performance Comparison 42
4.2.2 Per-Model Performance Comparison 42

UNIVERSITAS
GADJAH MADA
Preserving Functional Fidelity in Compressed Large Language Models Through Corrective and Adaptive Decomposition
Muchammad Daniyal Kautsar, Syukron Abu Ishaq Alfarozi, S.T., Ph.D.;Widyawan, S.T., M.Sc., Ph.D
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

4.3 Inference and Latency Analysis 43
4.4 Ablation Studies 44
4.4.1 Impact of the Corrective Module (F_{corr}) 44
4.4.2 Sensitivity to Corrective Module Rank (r_c) 45
4.4.3 Impact of Selective FFN Module Targeting 46
4.4.4 Analysis of Attention (QKV) Layer Compression 46
CHAPTER V Discussion of Results 48
5.1 Functional Residual Correction 48
5.1.1 Architectural Congruence with the Nature of Information Loss 48
5.1.2 Optimizing for Functional over Reconstruction Fidelity 49
5.2 Architectural Limitations and Methodological Scope 49
5.2.1 Justification of the Comparative Baseline Scope 49
5.2.2 The FFN-Centric Design and an Analysis of its Suboptimality for Attention Layers 50
5.3 Evaluating Research Hypotheses Against Empirical Results 51
CHAPTER VI Conclusion and Future Work 52
6.1 Conclusion 52
6.2 Future Work 53
REFERENCES 54
APPENDIX L-1
L.1 Formal Comparison of Data-Driven Compression Methodologies L-1
L.1.1 Notation L-1
L.1.2 Formalization of Whitening SVD Approaches L-1
L.1.3 Formalization of CALR L-2
L.1.4 Comparative Analysis of Data Influence L-3