Efficient Parameter Adaptation Methods for Large Language Models

A Unifying Framework

Wei Li

University of Birmingham & ZoyMed

Research Motivation and Challenges

Large Language Models demonstrate revolutionary performance but face critical deployment challenges that limit real-world applications. This research addresses three fundamental dimensions:

Parameter Scale

Massive parameter counts (7B-70B+) create prohibitive memory overhead and computational costs for deployment on resource-constrained devices.

Compression Methods

Traditional uniform pruning ignores layer-wise importance variations and requires extensive retraining, lacking theoretical foundations.

Theoretical Gap

Existing approaches rely on empirical hyperparameter tuning without principled frameworks for non-uniform, adaptive compression strategies.

Research Gap: Need for theoretically-grounded, training-free, and adaptive compression frameworks that exploit redundancy without sacrificing performance

Unified Framework Overview

A Cohesive Theoretical Framework Integrating Parameter-Efficient Adaptation Methods

Core Philosophy: Exploit redundancy and optimize parameter allocation without sacrificing performance

Adaptive Layer Sparsity (ALS) *Neurips 2024*

Linear programming for optimal layer-wise pruning ratios based on inter-layer redundancy analysis

Bayesian Knowledge Distillation *ACL·Findings 2025*

Fisher-informed distillation for compressed LLMs with logit dual-scaling and Bayesian optimization

MoE-SVD *ICML 2025*

Structured singular value decomposition for Mixture-of-Experts compression with V-matrix sharing

Delta Decompression (D²-MoE) *ICML 2025*

Base-delta weight decomposition addressing expert diversity through Fisher merging and structured pruning

Unifying Theme: Information-theoretic foundations combined with Bayesian optimization principles

Adaptive Layer Sparsity - Methodology

ALS Formulates Sparsity Allocation as Linear Programming with Redundancy Metrics

Redundancy Metric (RM)

Based on Centered Kernel Alignment (CKA):

$$RM(X_i, X_j) = \frac{\|X_i^T X_j\|^2}{\|X_i^T X_i\| \|X_j^T X_j\|}$$

Values: 0 = complete independence, 1 = complete redundancy

Optimization Objective

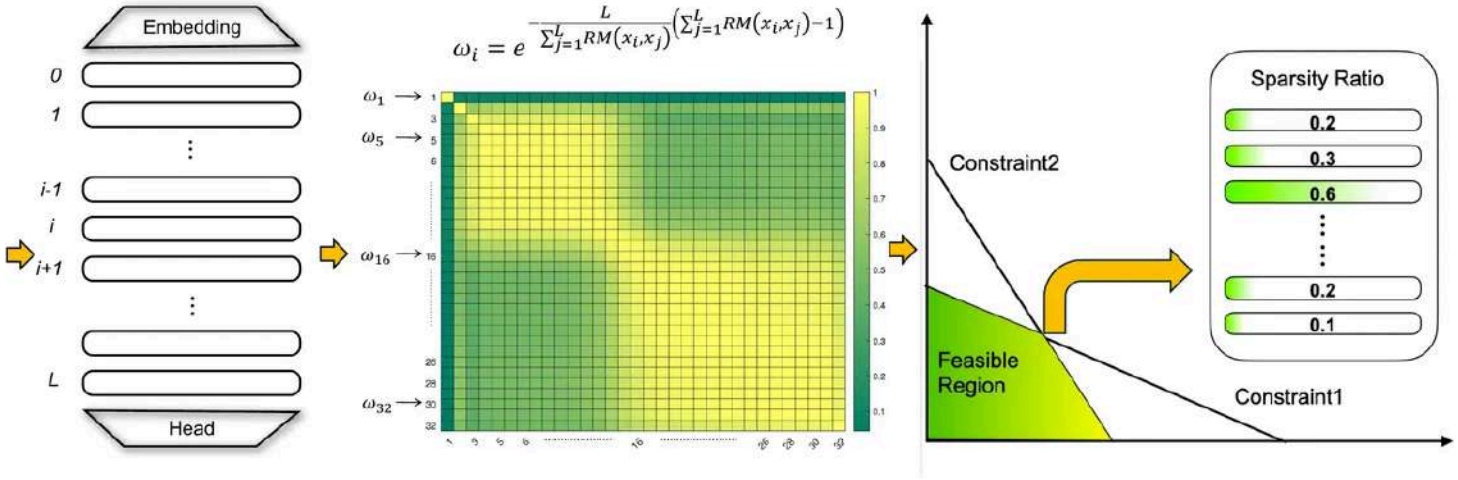
$$\max_q \sum_{i=1}^{L} \left(\frac{q_i}{L-i+1} \sum_{l=i}^{L} \omega_l \right)$$

where q_i is the sparsity rate for layer i

Constraint

$$\sum_{i=1}^{L} S^{(q_i)} \leq \text{Target Model Size}$$

Key Advantages



ALS Experimental Results

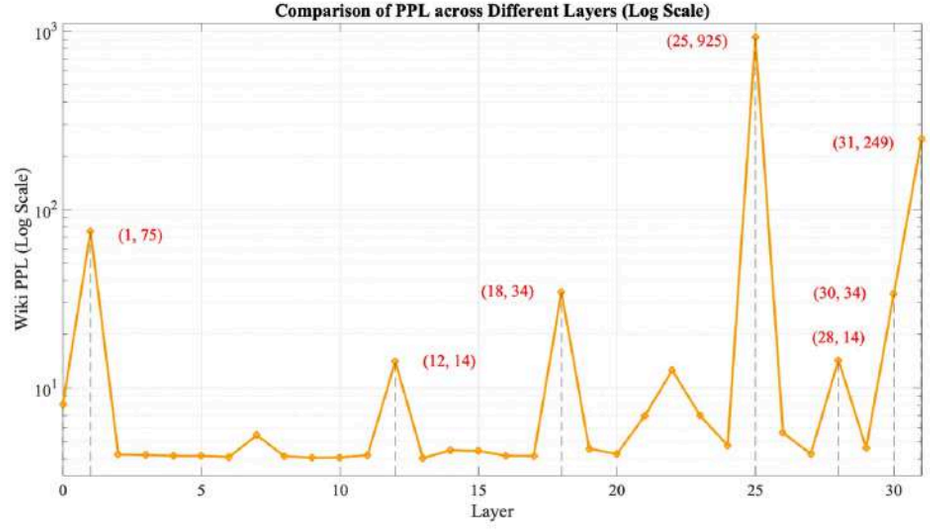
Superior Performance Over State-of-the-Art Methods at 50% Sparsity

WikiText-2 Perplexity Results

Model	Baseline	ALS	Improvement
LLaMA-V1 7B	Magnitude: 42.26	**16.80**	**60.3% ↓**
LLaMA-V1 13B	Magnitude: 43.61	**12.61**	**71.1% ↓**
LLaMA-V2 7B	Wanda: 11.21	**9.86**	**12.0% ↓**
LLaMA-V3 8B	Magnitude: 30.20	**13.21**	**56.3% ↓**

Zero-Shot Task Accuracy

Model	Baseline	ALS	Gain
LLaMA-V1 7B	Magnitude: 53.40	**56.28**	**+2.88**
LLaMA-V3 8B	Magnitude: 43.29	**57.43**	**+14.14**
OPT 6.7B	Wanda: 47.81	**47.89**	**+0.08**

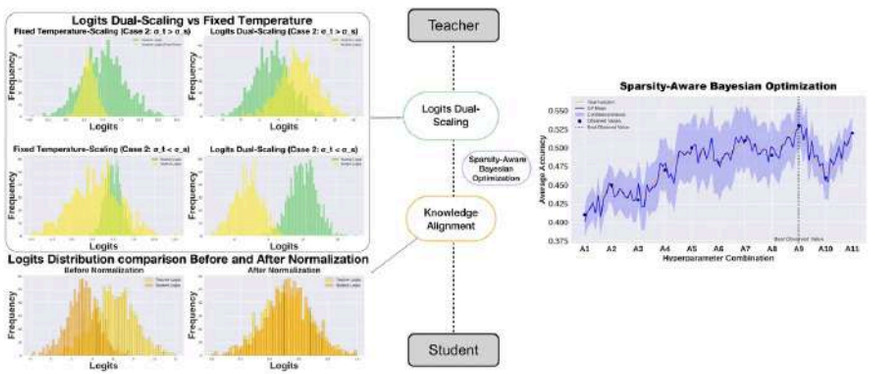


Baselines: Magnitude, SparseGPT, Wanda, OWL

Efficiency: 20 minutes for 70B model on single A100

Bayesian Knowledge Distillation (BayesKD)

Bridges the Logit Gap with Sparsity-Aware Bayesian Optimization



BayesKD Framework: Logits Dual-Scaling + Knowledge Alignment + Bayesian Optimization

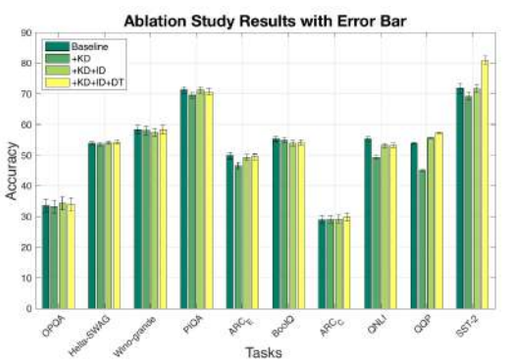
Performance Gains (Average Accuracy)

Teacher → Student	Baseline	BayesKD	Gain
LLaMA 7B → LLaMA 7B	Std KD: 38.60	**44.60**	**+6.0**
LLaMA3 13B → Tiny 1.1B	Std KD: 34.40	**34.00**	**+4.23***
LLaMA3 70B → LLaMA3 8B	Std KD: 60.47	**63.46**	**+2.99**
Qwen2 72B → Qwen2 7B	Std KD: 64.41	**68.46**	**+4.05**

*vs Sparse Model baseline (SABO optimization)

Three-Component Framework

1. **Logits Dual-Scaling:** Dynamically adjusts teacher/student logits based on standard deviations
2. **Knowledge Alignment:** Min-max normalization aligns intermediate representations
3. **Bayesian Optimization:** Sparsity-aware hyperparameter search



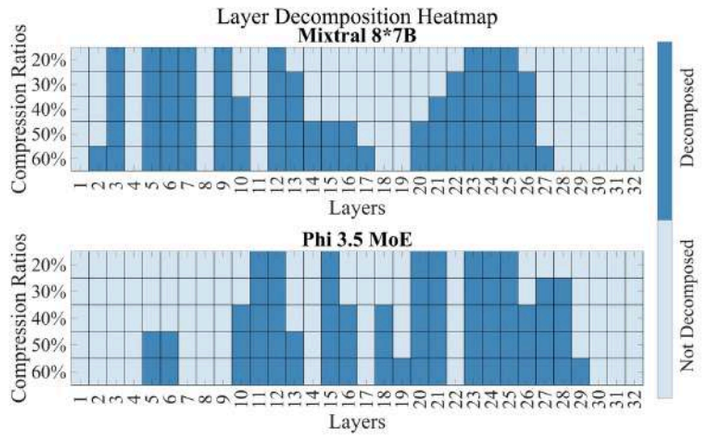
Ablation: Each component (KD, ID, DT) contributes to performance

MoE-SVD: Structured Compression

Selective Layer Decomposition Based on Sensitivity Analysis

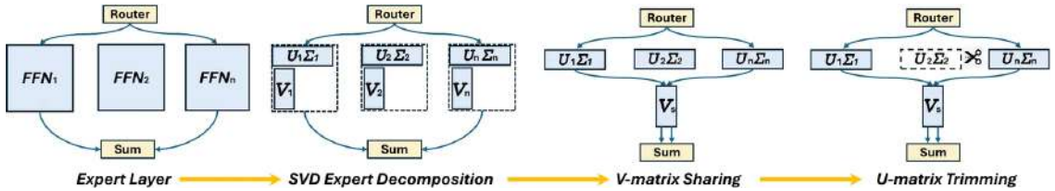
Selection Criteria

- **Sensitivity Metric:** Based on weight singular values and activation statistics
- **Layer-wise Decision:** Decompose only layers with low sensitivity to compression
- **Adaptive Strategy:** Different compression ratios per layer (20%-60%)



Layer-wise decomposition decisions for Mixtral-8×7B and Phi-3.5-MoE

MoE-SVD Pipeline



V-matrix sharing + U-matrix trimming pipeline

Key Techniques

1. **V-Matrix Sharing:** Shared across all experts
2. **U-Matrix Trimming:** Top-k selection
3. **Selective Decomposition:** Layer-specific

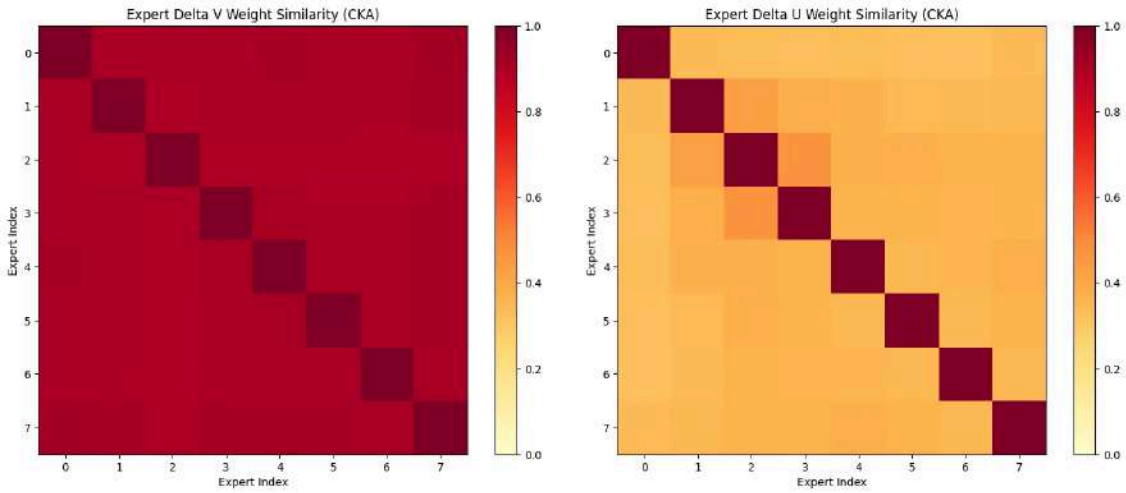
Experimental Results

Model	Compression	PPL Drop	Speedup	Baseline Comparison
Mixtral-8×7B	20%	2%	1.2×	Outperforms MC-SMoE
Mixtral-8×7B	40%	5%	1.5×	Best performance/compression trade-off
Phi-3.5-MoE	40%	5%	1.4×	Maintains 95% accuracy

Expert Redundancy and Delta Decompression (D²-MoE)

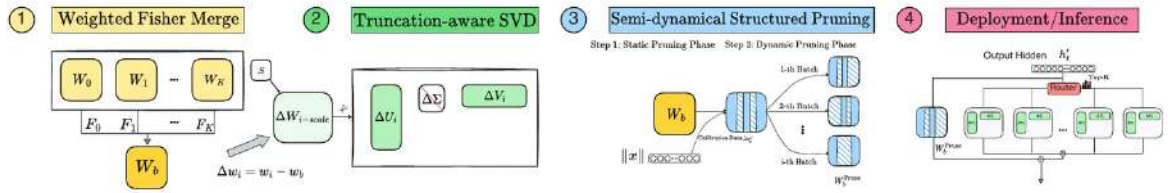
CKA Analysis Reveals Redundancy Patterns Enabling Delta-Based Compression

Expert Similarity Analysis



Key Observation: V-matrices show high redundancy (left), U-matrices show diversity (right)

D²-MoE Framework Pipeline



Four-stage pipeline: Weighted Fisher Merge → Truncation-aware SVD → Semi-dynamical Pruning → Deployment

D²-MoE Components

① **Weighted Fisher Merge:** Combines similar experts using Fisher information weighting (W_b = base weight)

Key Results

- **DeepSeek-MoE-16B:** 60% compression with minimal performance loss
- **Mixtral-8×7B:** Combined with MoE-SVD achieves aggressive compression

Comprehensive Performance Evaluation

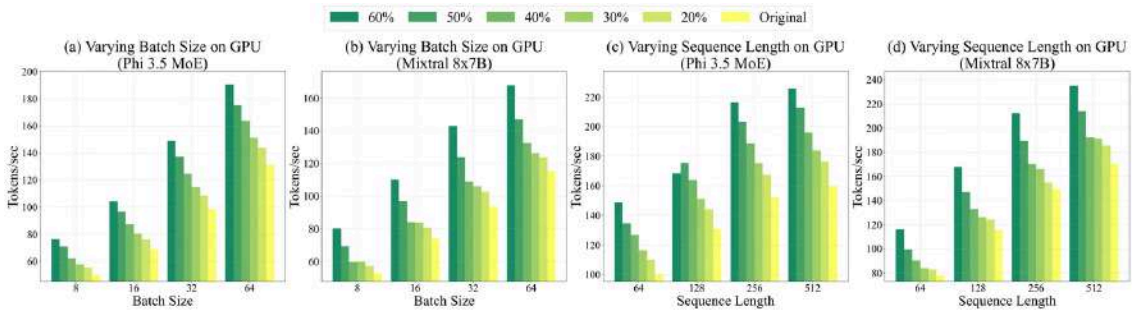
Unified Framework: Method-Specific Results Across Diverse Models

Method	Model Type	Key Achievement
Adaptive Layer Sparsity (ALS)		
ALS	LLaMA-2 7B-70B	**50% sparsity, 16% PPL improvement**
ALS	OPT 6.7B-13B	**20-70% sparsity range**
Bayesian Knowledge Distillation (BayesKD)		
BayesKD	LLaMA 13B → Tiny 1.1B	**+4.4% accuracy vs Standard KD**
BayesKD	Qwen-2 72B → 7B	**+2.6% vs LoRA fine-tuning**
MoE Compression (MoE-SVD + D²-MoE)		
MoE-SVD	Mixtral-8×7B	**40% compression, 5% drop, 1.5× speedup**
MoE-SVD	Phi-3.5-MoE	**40% compression, 5% drop**
D²-MoE	DeepSeek-MoE-16B	**60% compression, minimal loss**

Overall Framework Achievements

- ✓ **Parameter Reduction:** 40-80% across methods
- ✓ **Performance Retention:** 90-95% of original capability
- ✓ **Training Requirement:** Free or minimal fine-tuning
- ✓ **Inference Speedup:** 1.2-1.5× for MoE models

Inference Speed Comparison



Benchmarks: WikiText-2, OpenBookQA, HellaSwag, ARC, PIQA, BoolQ, SST-2, WinoGrande

Contributions and Future Directions

This Framework Establishes Theoretical Foundations for Next-Generation LLM Efficiency Research

Key Contributions

1. Theoretical Advancement

First linear programming formulation for LLM sparsity allocation with information-theoretic foundations

2. Practical Impact

Training-free methods enabling deployment on resource-constrained devices

3. Architectural Innovation

Specialized compression techniques for emerging MoE architectures

4. Empirical Validation

Extensive experiments across 6.7B-70B parameter models demonstrating scalability

Future Directions

- **Multimodal LLMs:** Extension to vision-language models and cross-modal compression
- **Combined Compression:** Integration with quantization for synergistic parameter reduction
- **Dynamic Adaptation:** Runtime sparsity adjustment based on input complexity
- **Federated Learning:** Application to distributed training scenarios with communication constraints

Publications

NeurIPS 2024 • ICML 2025 (×2) • ACL 2025

Thank You

Questions?

Zoy Technology Co., LtD

Reimagining precision medicine through world-class AI:
Empowering doctors, enabling patients.

SurgiFlow Cloud Platform

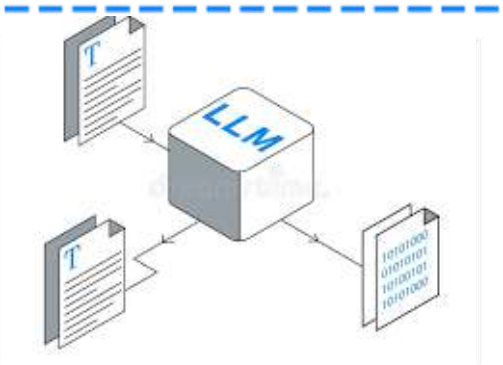
Hardware Services

Compact & light weight
Multi-models support
NFC wheat login

Encrypted Cloud

Secure Cloud Service

AI Analysis



Phase extraction
Dangerous operation
Critical operation
Surgical instruments detection

UI

Web