Novelty Assessment Report

Paper: KDP: Simplifying Representation Dynamics in Kernel Space

PDF URL: https://openreview.net/pdf?id=262LUKGdQn

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

This paper proposes Kernelized Dynamics Pruning (KDP), a novel layer pruning method from the perspective of simplifying representation dynamics within large language models (LLMs). Motivated by the high similarity between consecutive layer representations, we view the LLM's forward pass as a discrete-time dynamical system. We speculate that this phenomenon indicates the model's internal dynamics have entered a ``slow manifold'', which exhibits computational redundancy. Based on this insight, we project the representations into a kernel space where the complex, non-linear transformation between them is simplified to an approximately linear one. Then, a simple network learns the inverse kernel transformation, thereby enabling the pruning of the entire layer block. Both theoretical analysis and extensive experiments validate the effectiveness of KDP, demonstrating its superiority over existing pruning baselines. Code is available at https://anonymous.4open.science/r/draft-123abc.

Core Task Landscape

This paper addresses: **Simplifying Representation Dynamics in Large Language Models through Layer Pruning**

A total of **49 papers** were analyzed and organized into a taxonomy with **32 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Layer Importance Analysis and Measurement**
- **Layer Removal Strategies**
- **Representation Dynamics and Theoretical Foundations**
- **Compensation Mechanisms for Layer Removal**
- **Intra-Layer Pruning Approaches**
- **Task-Specific and Application-Oriented Pruning**
- **Hybrid and Multi-Strategy Pruning Frameworks**
- **Architectural Variants and Layer Design**
- **Specialized Pruning Contexts and Constraints**
- **General Structural Pruning Frameworks**

Complete Taxonomy Tree

- Simplifying Representation Dynamics in Large Language Models through Layer Pruning Survey Taxonomy
- Layer Importance Analysis and Measurement
  - Activation-Based Layer Importance Metrics (2 papers)
  - [4] Layer Importance and Hallucination Analysis in Large Language Models via Enhanced Activation Variance-Sparsity (Song Zichen, 2024) View paper
  - [16] Adaptive Layer Sparsity for Large Language Models via Activation Correlation Assessment (Mark Lee, 2024) View paper
  - Similarity-Based Layer Importance Metrics (3 papers)
  - [2] The unreasonable ineffectiveness of the deeper layers (Gromov, 2024) View paper
  - [18] Change Is the Only Constant: Dynamic LLM Slicing based on Layer Redundancy (Razvan-Gabriel Dumitru, 2024) View paper
  - [45] ShortGPT: Layers in Large Language Models are More Redundant Than You Expect (Men Xin, 2024) View paper
  - Shapley Value and Attribution-Based Importance (1 papers)
  - [14] Investigating layer importance in large language models (Zhang Yang, 2024) View paper
  - Topological and Geometric Layer Analysis (1 papers)
  - [20] Persistent Topological Features in Large Language Models (Viswanathan, 2024) View paper
- Layer Removal Strategies
  - Uniform and Block-Based Layer Removal (3 papers)
  - [1] Why lift so heavy? slimming large language models by cutting off the layers (Yuan, 2025) View paper
  - [6] Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks (Song Ji-won, 2024) View paper
  - [11] Streamlining redundant layers to compress large language models (Chen, 2024) View paper
  - Non-Uniform and Layerwise-Adaptive Removal (2 papers)
  - [3] DLP: Dynamic Layerwise Pruning in Large Language Models (Chen Yu-li, 2025) View paper
  - [21] High-Layer Attention Pruning with Rescaling (Liu Song-tao, 2025) View paper
  - Dynamic and Adaptive Layer Removal (5 papers)
  - [23] TELL-TALE: Task Efficient LLMs with Task Aware Layer Elimination (Sharma, 2025) View paper
  - [25] Dynamic layer selection in decoder-only transformers (Theodore Glavas, 2024) View paper
  - [28] Dynamic Layer Skipping for Large Language Models on Natural Language Understanding Tasks and Machine Translation Using Reinforcement Learning (Wei Xu, 2024) View paper

    - [31] Hopscotch: Discovering and Skipping Redundancies in Language Models (Eyceoz, 2025) View paper
    - [33] SkipGPT: Dynamic Layer Pruning Reinvented with Token Awareness and Module Decoupling (Ye Fanghua, 2025) View paper
    - Layer Merging and Stitching (1 papers)
    - [35] GPTailor: Large Language Model Pruning Through Layer Cutting and Stitching (Su, 2025) View paper
- Representation Dynamics and Theoretical Foundations
    - Kernel Space and Dynamical Systems Perspectives ★ (1 papers)
    - [0] KDP: Simplifying Representation Dynamics in Kernel Space (Anon et al., 2026) View paper
    - Robustness and Stages of Inference (2 papers)
    - [10] The remarkable robustness of llms: Stages of inference? (Lad, 2024) View paper
    - [27] Demystifying the roles of llm layers in retrieval, knowledge, and reasoning (Song Xinyuan, 2025) View paper
    - Layer Functionality and Knowledge Localization (1 papers)
    - [13] On the Fragility of Latent Knowledge: Layer-wise Influence under Unlearning in Large Language Model (J Zhu, 2025) View paper
- Compensation Mechanisms for Layer Removal
    - Magnitude and Scaling Compensation (2 papers)
    - [24] Prune&Comp: Free Lunch for Layer-Pruned LLMs via Iterative Pruning with Magnitude Compensation (Chen Xin-rui, 2025) View paper
    - Gating and Identity Mapping Mechanisms (1 papers)
    - [47] Learning Identity Mappings with Residual Gates (Savarese, 2016) View paper
- Intra-Layer Pruning Approaches
    - Attention Head and Component Pruning (1 papers)
    - [15] LLM-BIP: Structured Pruning for Large Language Models with Block-Wise Forward Importance Propagation (Wu, 2024) View paper
    - Structured Sparsity and N:M Patterns (1 papers)
    - [12] Accelerating LLM Inference with Flexible N:M Sparsity via A Fully Digital Compute-in-Memory Accelerator (Kundu, 2025) View paper
    - Low-Rank Approximation and Factorization (2 papers)
    - [34] GRASP: Replace Redundant Layers with Adaptive Singular Parameters for Efficient Model Compression (Kainan Liu, 2025) View paper
    - [37] Adaptive Rank Pruning: Dynamic Low-Rank Model Merging and Compression for Efficient AI Deployment (Ieee M. VEDHANTH Student Member, 2025) View paper
    - Token and KV Cache Pruning (3 papers)
    - [8] SlimInfer: Accelerating Long-Context LLM Inference via Dynamic Token Pruning (Yang Rubing, 2025) View paper
    - [17] Lethe: Layer- and Time-Adaptive KV Cache Pruning for Reasoning-Intensive LLM Serving (Hui Zeng, 2025) View paper
    - [39] OBCache: Optimal Brain KV Cache Pruning for Efficient Long-Context LLM Inference (Gu, 2025) View paper
- Task-Specific and Application-Oriented Pruning
    - Task-Aware Layer Selection (1 papers)
    - [41] How Many Parameters Does Your Task Really Need? Task Specific Pruning with LLM-Sieve (Reda, 2025) View paper
    - Prompt-Based and Fine-Tuning Contexts (1 papers)
    - [26] Efficient Federated Fine-Tuning of Large Language Models with Layer Dropout (Wang Shi-long, 2025) View paper
    - Domain-Specific Applications (3 papers)
    - [19] SLMRec: Distilling Large Language Models into Small for Sequential Recommendation (Xu Wujiang, 2024) View paper
    - [22] Efficient contextualized representation: Language model pruning for sequence labeling (Liyuan Liu, 2018) View paper
    - [30] PUMA: Layer-Pruned Language Model for Efficient Unified Multimodal Retrieval with Modality-Adaptive Learning (Yibo Lyu, 2025) View paper
- Hybrid and Multi-Strategy Pruning Frameworks
    - Layer Removal with Intra-Layer Pruning (1 papers)
    - [48] Rethinking Layer Removal: A Hybrid Pruning Framework Combining Layer Removal and Singular Value Selection for Efficient LLM Compression (Kainan Liu, n.d.) View paper
    - Iterative and Multi-Stage Pruning (1 papers)
    - [36] Iterative Layer-wise Distillation for Efficient Compression of Large Language Models (Tikhomirov, 2025) View paper
    - Partition-Based and Similarity-Guided Pruning (1 papers)
    - [42] SGLP: A Similarity Guided Fast Layer Partition Pruning for Compressing Large Deep Models (Li YuQi, 2024) View paper
- Architectural Variants and Layer Design
    - Layer-Wise Scaling and Width Variants (1 papers)
    - [29] Crown, Frame, Reverse: Layer-Wise Scaling Variants for LLM Pre-Training (Andrei Baroian, 2025) View paper
    - Structural Permutation and Modulation (2 papers)
    - [7] Structural permutation layers: An unprecedented approach for modulating internal representations in large language models (Watson, 2025) View paper
    - [49] Interpretable Structural Drift Modulation for Large Language Model Transformer Pathways through Recursive Signal Recomposition (O Finch, n.d.) View paper
- Specialized Pruning Contexts and Constraints
    - Real-Time and Embedded Systems Pruning (1 papers)
    - [44] NetCut: Real-Time DNN Inference Using Layer Removal (Mehrshad Zandigohar, 2021) View paper
    - Contextual Compression and Encoding (1 papers)
    - [9] Contextual compression encoding for large language models: A novel framework for multi-layered parameter space pruning (Barnaby Schmitt, 2025) View paper
    - Truthfulness and Safety-Preserving Pruning (1 papers)
    - [32] Pruning Weights but Not Truth: Safeguarding Truthfulness While Pruning LLMs (Fu Yao, 2025) View paper
    - Long-Context and Reasoning-Intensive Scenarios (1 papers)
    - [40] When Fewer Layers Break More Chains: Layer Pruning Harms Test-Time Scaling in LLMs (Wang Keyu, 2025) View paper
- General Structural Pruning Frameworks
    - Task-Agnostic Structural Pruning (1 papers)

- [5] Llm-pruner: On the structural pruning of large language models (Ma, 2023) View paper
- Interpretable and Fine-Grained Pruning (1 papers)
- [43] FinerCut: Finer-grained Interpretable Layer Pruning for Large Language Models (Zhang Yang, 2024) View paper
- Reweighted and Proximal Pruning (1 papers)
- [46] Reweighted Proximal Pruning for Large-Scale Language Representation (Guo, 2022) View paper

Narrative

Core task: Simplifying representation dynamics in large language models through layer pruning. The field has organized itself around several complementary perspectives. At the highest level, researchers distinguish between methods that analyze which layers matter (Layer Importance Analysis and Measurement), strategies for actually removing layers (Layer Removal Strategies), and theoretical work examining how representations evolve across depth (Representation Dynamics and Theoretical Foundations). Parallel branches address compensation mechanisms that restore performance after pruning, intra-layer pruning that targets weights or attention heads rather than entire layers, and task-specific approaches tailored to particular applications. Hybrid frameworks combine multiple pruning strategies, while other branches explore architectural variants and specialized contexts such as federated learning or constrained deployment. Representative works like Slimming LLMs[1] and ShortGPT Layer Redundancy[45] illustrate how layer removal strategies operate in practice, whereas Investigating Layer Importance[14] and Layer Importance Hallucination[4] exemplify efforts to measure and understand layer contributions.

A central tension runs through the literature: some studies argue that deeper layers become less effective or even redundant (Deeper Layers Ineffectiveness[2]), while others find that high layers play critical roles in specific tasks (High Layer Attention[21]). Dynamic approaches like Dynamic Layerwise Pruning[3] and Dynamic Layer Selection[25] attempt to reconcile these views by adapting pruning decisions to input or task context. The original paper, KDP Kernel Space[0], sits within the theoretical foundations branch, offering a dynamical systems perspective on how layer transformations evolve. This contrasts with more empirical measurement studies like Investigating Layer Importance[14] and complements structural methods such as LLM Pruner[5], which focus on dependency-aware removal. By framing layer dynamics through kernel space analysis, KDP Kernel Space[0] provides a mathematical lens that bridges abstract representation theory with practical pruning decisions, addressing why certain layers can be simplified without degrading model behavior.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on theoretical frameworks that model layer transformations through dynamical systems or kernel space projections, providing mathematical abstractions for understanding representation evolution. The sibling subtopics examine complementary aspects: one investigates what knowledge different layers encode and their functional specialization, while the other empirically tests how models respond to layer interventions and identifies distinct computational stages. Together, these subtopics span theoretical modeling, functional analysis, and empirical robustness testing of layer dynamics.

Similarities: - All three subtopics analyze layer-wise behavior and transformations in large language models - Each examines how representations evolve or change across network depth - All are relevant to understanding which layers can be removed or modified (layer pruning context) - Each subtopic excludes overlapping concerns with the others, suggesting they form complementary perspectives on the same phenomenon

Differences: - The original leaf takes a theoretical/mathematical modeling approach (dynamical systems, kernel spaces), while siblings focus on empirical analysis - Layer Functionality examines what is encoded (knowledge localization), Robustness examines resilience to interventions, while the original leaf examines how transformations occur mathematically - The original leaf explicitly excludes empirical robustness studies, which are the focus of the Robustness sibling - Layer Functionality emphasizes static functional roles and knowledge distribution, whereas the original leaf emphasizes dynamic transformation processes

Suggested Search Directions: - Connections between dynamical systems theory and empirically observed inference stages - How kernel space projections relate to layer-specific knowledge encoding patterns - Mathematical frameworks that predict which layers are robust to removal based on dynamical properties - Unified models combining theoretical dynamics with functional localization findings

Sibling Subtopics

- **Layer Functionality and Knowledge Localization** (leaves: 1, papers: 1)
- Scope: Research on how different layers encode knowledge and their functional roles.
- Exclude: Excludes pruning methods; see Layer Removal Strategies or Layer Importance Analysis.
- **Robustness and Stages of Inference** (leaves: 1, papers: 2)
- Scope: Studies examining model robustness to layer interventions and identifying inference stages.
- Exclude: Excludes theoretical dynamical models; see Kernel Space and Dynamical Systems Perspectives.

Contributions Analysis

Overall novelty summary. The paper introduces Kernelized Dynamics Pruning (KDP), which frames layer pruning through a dynamical systems lens by projecting representations into kernel space to linearize transformations. Within the taxonomy, it occupies the 'Kernel Space and Dynamical Systems Perspectives' leaf under 'Representation Dynamics and Theoretical Foundations'. Notably, this leaf contains only the original paper itself—no sibling papers exist in this specific category. This positioning suggests the work explores a relatively sparse theoretical direction, distinct from the more populated empirical branches like 'Uniform and Block-Based Layer Removal' or 'Similarity-Based Layer Importance Metrics'.

The taxonomy reveals that neighboring leaves focus on empirical robustness studies ('Robustness and Stages of Inference') and knowledge localization ('Layer Functionality and Knowledge Localization'), while sibling branches address practical removal strategies and compensation mechanisms. The 'Representation Dynamics and Theoretical Foundations' parent branch itself is less crowded than 'Layer Removal Strategies', which contains multiple subtopics with numerous papers. KDP's kernel-based formulation diverges from activation-based or similarity-based importance metrics, instead offering a mathematical framework that connects to but does not directly overlap with empirical layer removal methods like ShortGPT or Slimming LLMs.

Across three contributions, the analysis examined 17 candidate papers total, with no clear refutations identified. The core KDP method examined 5 candidates with 0 refutable matches; the theoretical error bound examined 10 candidates with 0 refutations; and the geometric embedding reformulation examined 2 candidates with 0 refutations. Among the limited search scope of top-K semantic matches, no prior work appears to provide the same kernel-space linearization approach combined with learned inverse transformations for layer pruning. The theoretical contributions, particularly the error bound and geometric reformulation, show no overlapping prior work within the examined candidates.

Based on the limited literature search of 17 candidates, the work appears to occupy a novel theoretical niche within layer pruning research. The absence of sibling papers in its taxonomy leaf and the lack of refutable prior work among examined candidates suggest

distinctiveness, though the search scope does not cover the entire field. The kernel-based dynamical systems perspective represents a less-explored angle compared to the more populated empirical and heuristic pruning branches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Kernelized Dynamics Pruning (KDP) method

Description: The authors introduce KDP, a layer pruning approach that projects LLM representations into a kernel space where complex non-linear transformations between layers are simplified to approximately linear ones, enabling entire layer blocks to be pruned while maintaining performance.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. SkipGPT: Dynamic Layer Pruning Reinvented with Token Awareness and Module Decoupling

URL: View paper

Brief Assessment

SkipGPT Token Awareness[33] focuses on dynamic, token-aware layer pruning with module decoupling (attention vs. MLP), rather than kernel space representation simplification for layer pruning.

---

2. Contextual compression encoding for large language models: A novel framework for multi-layered parameter space pruning

URL: View paper

Brief Assessment

Contextual Compression Encoding[9] focuses on structured parameter space pruning through contextual similarity metrics and multi-layer encoding, not on kernel space transformations or representation dynamics simplification as proposed in KDP.

---

3. Change Is the Only Constant: Dynamic LLM Slicing based on Layer Redundancy

URL: View paper

Brief Assessment

Dynamic LLM Slicing[18] focuses on dynamic layer-specific pruning using a layer redundancy score based on cosine similarity, while KDP projects representations into kernel space for linearization. These are fundamentally different technical approaches to layer pruning.

---

4. Emergent Crystallographic Inference Fields in Large Language Models: A Nonlinear Inductive Geometry for Probabilistic Decay Alignment

URL: View paper

Brief Assessment

Crystallographic Inference Fields[51] focuses on crystallographic inference and probabilistic decay alignment in LLMs, not on layer pruning methods using kernel space representation dynamics simplification as proposed in KDP.

---

5. How can representation dimension dominate structurally pruned LLMs?

URL: View paper

Brief Assessment

Representation Dimension Pruning[50] focuses on how representation dimension affects structured pruning outcomes through analytical relations, rather than proposing a kernel-space linearization method for layer pruning like KDP.

Contribution 2: Theoretical error bound for kernel linearization

Description: The authors establish Theorem 1 providing an error bound for approximating multi-layer representations with linear transformations in kernel space, and Theorem 2 demonstrating that kernel space exhibits superior fitting capacity compared to the original representation space.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Neural Tangent Kernel: Convergence and Generalization in Neural Networks

URL: View paper

Brief Assessment

Neural Tangent Kernel[54] focuses on describing neural network training dynamics through a fixed kernel in the infinite-width limit, not on error bounds for approximating multi-layer representations with linear transformations in kernel space for model compression purposes.

---

2. Data-Efficient Kernel Methods for Learning Differential Equations and Their Solution Operators: Algorithms and Error Analysis

URL: View paper

Brief Assessment

Kernel Methods PDEs[57] focuses on learning differential equations using kernel methods with error bounds for approximating PDEs and their solution operators. The original paper addresses kernel linearization of neural network layer transformations in LLMs, which is a fundamentally different application domain and mathematical framework.

---

3. Solving Roughly Forced Nonlinear PDEs via Misspecified Kernel Methods and Neural Networks

URL: View paper

Brief Assessment

Misspecified Kernel Methods[59] addresses solving roughly forced nonlinear PDEs using kernel methods with misspecified kernels, focusing on error bounds for approximating PDE solutions in RKHS. The original paper's contribution concerns error bounds for approximating multi-layer neural network representations with linear transformations in kernel space for LLM pruning, which is a fundamentally different problem domain and technical setting.

---

4. An introduction to kernel-based learning algorithms

URL: View paper

Brief Assessment

Kernel Based Learning[55] focuses on kernel methods for SVMs, kernel PCA, and kernel Fisher discriminant, not on error bounds for approximating multi-layer neural network representations with linear transformations in kernel space as proposed in the original paper.

---

5. A theoretical analysis of the test error of finite-rank kernel ridge regression

URL: View paper

Brief Assessment

Finite Rank Kernel[61] focuses on kernel ridge regression (KRR) for statistical learning with finite-rank kernels, deriving test error bounds. The original paper addresses kernel space linearization for layer pruning in LLMs, a fundamentally different application domain and technical problem.

---

6. Neural hilbert ladders: Multi-layer neural networks in function space

URL: View paper

Brief Assessment

Neural Hilbert Ladders[56] focuses on characterizing function spaces of multi-layer neural networks through hierarchical RKHSs and provides approximation bounds for functions in these spaces, not on error bounds for kernel space linearization in layer transformations of LLMs.

---

7. Linearized two-layers neural networks in high dimension

URL: View paper

Brief Assessment

Linearized Two Layers[52] analyzes random feature and neural tangent kernel models on spherical data, proving polynomial approximation bounds. The original paper establishes error bounds for kernel space linearization in transformer layer dynamics—a fundamentally different architectural context and problem setting.

---

8. Soft: Softmax-free transformer with linear complexity

URL: View paper

Brief Assessment

Softmax Free Transformer[58] focuses on linearizing self-attention in vision transformers using Gaussian kernels and Nyström decomposition, not on kernel space linearization for layer pruning in LLMs. The theoretical frameworks address fundamentally different problems.

---

9. Spectrum dependent learning curves in kernel regression and wide neural networks

URL: View paper

Brief Assessment

Spectrum Learning Curves[53] focuses on generalization error bounds for kernel regression and neural networks in the infinite-width limit, not on error bounds for approximating multi-layer transformations with linear operations in kernel space as proposed in the original paper.

---

10. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit

URL: View paper

Brief Assessment

Mean Field Theory[60] focuses on approximating neural network training dynamics through mean-field PDEs in probability distribution space, not on kernel space linearization for layer pruning in LLMs. The theoretical framework addresses gradient descent convergence in two-layer networks rather than error bounds for transforming multi-layer representations via kernel projections.

Contribution 3: Reformulation of layer pruning as geometric embedding search

Description: The authors reframe the layer pruning problem as finding an optimal geometric viewpoint in a Reproducing Kernel Hilbert Space where the inherent simplicity of complex dynamics can be revealed, rather than merely constructing smaller sub-networks.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Modeling and analyzing neural networks using reproducing kernel Hilbert space algorithm

URL: View paper

Brief Assessment

Reproducing Kernel RKHS[62] focuses on general neural network modeling and analysis using RKHS theory, not on layer pruning or geometric embedding optimization for model compression.

---

2. An Analysis of Reinforcement Learning in High Dimensions with Kernel and Neural Network Approximation

URL: View paper

Brief Assessment

Reinforcement Learning Kernels[63] focuses on kernel-based function approximation for Q-learning in RL with high-dimensional state spaces, not on layer pruning or geometric embedding optimization in neural network compression contexts.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] KDP: Simplifying Representation Dynamics in Kernel Space View paper
- [1] Why lift so heavy? slimming large language models by cutting off the layers View paper
- [2] The unreasonable ineffectiveness of the deeper layers View paper
- [3] DLP: Dynamic Layerwise Pruning in Large Language Models View paper
- [4] Layer Importance and Hallucination Analysis in Large Language Models via Enhanced Activation Variance-Sparsity View paper
- [5] Llm-pruner: On the structural pruning of large language models View paper
- [6] Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks View paper
- [7] Structural permutation layers: An unprecedented approach for modulating internal representations in large language models View paper

- [8] SlimInfer: Accelerating Long-Context LLM Inference via Dynamic Token Pruning View paper
- [9] Contextual compression encoding for large language models: A novel framework for multi-layered parameter space pruning View paper
- [10] The remarkable robustness of llms: Stages of inference? View paper
- [11] Streamlining redundant layers to compress large language models View paper
- [12] Accelerating LLM Inference with Flexible N:M Sparsity via A Fully Digital Compute-in-Memory Accelerator View paper
- [13] On the Fragility of Latent Knowledge: Layer-wise Influence under Unlearning in Large Language Model View paper
- [14] Investigating layer importance in large language models View paper
- [15] LLM-BIP: Structured Pruning for Large Language Models with Block-Wise Forward Importance Propagation View paper
- [16] Adaptive Layer Sparsity for Large Language Models via Activation Correlation Assessment View paper
- [17] Lethe: Layer- and Time-Adaptive KV Cache Pruning for Reasoning-Intensive LLM Serving View paper
- [18] Change Is the Only Constant: Dynamic LLM Slicing based on Layer Redundancy View paper
- [19] SLMRec: Distilling Large Language Models into Small for Sequential Recommendation View paper
- [20] Persistent Topological Features in Large Language Models View paper
- [21] High-Layer Attention Pruning with Rescaling View paper
- [22] Efficient contextualized representation: Language model pruning for sequence labeling View paper
- [23] TELL-TALE: Task Efficient LLMs with Task Aware Layer Elimination View paper
- [24] Prune&Comp: Free Lunch for Layer-Pruned LLMs via Iterative Pruning with Magnitude Compensation View paper
- [25] Dynamic layer selection in decoder-only transformers View paper
- [26] Efficient Federated Fine-Tuning of Large Language Models with Layer Dropout View paper
- [27] Demystifying the roles of llm layers in retrieval, knowledge, and reasoning View paper
- [28] Dynamic Layer Skipping for Large Language Models on Natural Language Understanding Tasks and Machine Translation Using Reinforcement Learning View paper
- [29] Crown, Frame, Reverse: Layer-Wise Scaling Variants for LLM Pre-Training View paper
- [30] PUMA: Layer-Pruned Language Model for Efficient Unified Multimodal Retrieval with Modality-Adaptive Learning View paper
- [31] Hopscotch: Discovering and Skipping Redundancies in Language Models View paper
- [32] Pruning Weights but Not Truth: Safeguarding Truthfulness While Pruning LLMs View paper
- [33] SkipGPT: Dynamic Layer Pruning Reinvented with Token Awareness and Module Decoupling View paper
- [34] GRASP: Replace Redundant Layers with Adaptive Singular Parameters for Efficient Model Compression View paper
- [35] GPTailor: Large Language Model Pruning Through Layer Cutting and Stitching View paper
- [36] Iterative Layer-wise Distillation for Efficient Compression of Large Language Models View paper
- [37] Adaptive Rank Pruning: Dynamic Low-Rank Model Merging and Compression for Efficient AI Deployment View paper
- [38] Prune&Comp: Free Lunch for Layer-Pruned LLMs via Iterative Pruning with Magnitude Compensation View paper
- [39] OBCache: Optimal Brain KV Cache Pruning for Efficient Long-Context LLM Inference View paper
- [40] When Fewer Layers Break More Chains: Layer Pruning Harms Test-Time Scaling in LLMs View paper
- [41] How Many Parameters Does Your Task Really Need? Task Specific Pruning with LLM-Sieve View paper
- [42] SGLP: A Similarity Guided Fast Layer Partition Pruning for Compressing Large Deep Models View paper
- [43] FinerCut: Finer-grained Interpretable Layer Pruning for Large Language Models View paper
- [44] NetCut: Real-Time DNN Inference Using Layer Removal View paper
- [45] ShortGPT: Layers in Large Language Models are More Redundant Than You Expect View paper
- [46] Reweighted Proximal Pruning for Large-Scale Language Representation View paper
- [47] Learning Identity Mappings with Residual Gates View paper
- [48] Rethinking Layer Removal: A Hybrid Pruning Framework Combining Layer Removal and Singular Value Selection for Efficient LLM Compression View paper
- [49] Interpretable Structural Drift Modulation for Large Language Model Transformer Pathways through Recursive Signal Recomposition View paper
- [50] How can representation dimension dominate structurally pruned LLMs? View paper
- [51] Emergent Crystallographic Inference Fields in Large Language Models: A Nonlinear Inductive Geometry for Probabilistic Decay Alignment View paper
- [52] Linearized two-layers neural networks in high dimension View paper
- [53] Spectrum dependent learning curves in kernel regression and wide neural networks View paper
- [54] Neural Tangent Kernel: Convergence and Generalization in Neural Networks View paper
- [55] An introduction to kernel-based learning algorithms View paper
- [56] Neural hilbert ladders: Multi-layer neural networks in function space View paper
- [57] Data-Efficient Kernel Methods for Learning Differential Equations and Their Solution Operators: Algorithms and Error Analysis View paper
- [58] Soft: Softmax-free transformer with linear complexity View paper
- [59] Solving Roughly Forced Nonlinear PDEs via Misspecified Kernel Methods and Neural Networks View paper
- [60] Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit View paper
- [61] A theoretical analysis of the test error of finite-rank kernel ridge regression View paper
- [62] Modeling and analyzing neural networks using reproducing kernel Hilbert space algorithm View paper
- [63] An Analysis of Reinforcement Learning in High Dimensions with Kernel and Neural Network Approximation View paper