Novelty Assessment Report

Paper: Sequences of Logits Reveal the Low Rank Structure of Language Models

PDF URL: https://openreview.net/pdf?id=gdZ6J5hZzF

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

A major problem in the study of large language models is to understand their inherent low-dimensional structure. We introduce an approach to study the low-dimensional structure of language models at a model-agnostic level: as sequential probabilistic models. We first empirically demonstrate that a wide range of modern language models exhibit low-rank structure: in particular, matrices built from the model's logits for varying sets of prompts and responses have low approximate rank. We then show that this low-rank structure can be leveraged for generation --- in particular, we can generate a response to a target prompt using a linear combination of the model's outputs on unrelated, or even nonsensical prompts.

On the theoretical front, we observe that studying the approximate rank of language models in the sense discussed above yields a simple universal abstraction whose theoretical predictions parallel our experiments. We then analyze the representation power of the abstraction and give provable learning guarantees.

Core Task Landscape

This paper addresses: **Understanding Low-Dimensional Structure of Language Models Through Logit Matrices**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

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

- **Theoretical Foundations and Empirical Analysis of Low-Rank Structure**
- **Low-Rank Adaptation Methods for Parameter-Efficient Fine-Tuning**
- **Model Compression via Matrix and Tensor Factorization**
- **Quantization-Aware Low-Rank Methods**
- **Dimensionality Reduction for Embeddings and Representations**
- **General Dimensionality Reduction Methods and Surveys**
- **Low-Rank Structure in Specialized Applications**

Complete Taxonomy Tree

- Understanding Low-Dimensional Structure of Language Models Through Logit Matrices Survey Taxonomy
- Theoretical Foundations and Empirical Analysis of Low-Rank Structure
  - Intrinsic Dimensionality and Rank Analysis ★ (3 papers)
  - [0] Sequences of Logits Reveal the Low Rank Structure of Language Models (Anon et al., 2026) View paper
  - [1] Intrinsic dimensionality explains the effectiveness of language model fine-tuning (Aghajanyan, 2021) View paper
  - [6] Bridging the dimensional chasm: Uncover layer-wise dimensional reduction in transformers through token correlation (Song Zhuo-yang, 2025) View paper
  - Geometric and Algebraic Frameworks for Language Model Structure (2 papers)
  - [9] Implicit Geometry of Next-token Prediction: From Language Sparsity Patterns to Model Representations (Zhao, 2024) View paper
  - [48] Geometry is all you need: A unified taxonomy of matrix and tensor factorization for compression of generative language models (Xu, 2024) View paper
  - Representation Analysis and Feature Geometry (5 papers)
  - [4] Rediscovering the latent dimensions of personality with large language models as trait descriptors (Joseph Suh, 2024) View paper
  - [5] Analysis of argument structure constructions in the large language model BERT (Pegah Ramezani, 2025) View paper
  - [20] Analysis of argument structure constructions in a deep recurrent language model (Pegah Ramezani, 2025) View paper
  - [25] Not all language model features are one-dimensionally linear (Engels, 2024) View paper
  - [43] Talking Heads: Understanding Inter-layer Communication in Transformer Language Models (Carsten Eickhoff, 2024) View paper
- Low-Rank Adaptation Methods for Parameter-Efficient Fine-Tuning
  - Core Low-Rank Adaptation Frameworks (3 papers)
  - [10] LoRA: Low-Rank Adaptation of Large Language Models (Hu, 2021) View paper
  - [21] Ensemble of low-rank adapters for large language model fine-tuning (X Wang, 2023) View paper
  - [32] Large Language Model Fine-tuning with Low-Rank Adaptation: A Performance Exploration (Bagus Hanindhito, 2025) View paper
  - Mixture-of-Experts and Dynamic Low-Rank Adaptation (4 papers)
  - [19] X-LoRA: Mixture of Low-Rank Adapter Experts, a Flexible Framework for Large Language Models with Applications in Protein Mechanics and Design (Eric L. Buehler, 2024) View paper

    - [28] AdaMoLE: Fine-Tuning Large Language Models with Adaptive Mixture of Low-Rank Adaptation Experts (Liu Ze-fang, 2024) View paper
    - [29] Multiple choice learning of low rank adapters for language modeling (Fontaine, 2025) View paper
    - [45] QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large Language Model Tuning (Chen, 2024) View paper
    - Parameter Sharing and Structured Low-Rank Adaptation (2 papers)
    - [24] ShareLoRA: Parameter Efficient and Robust Large Language Model Fine-tuning via Shared Low-Rank Adaptation (Song, 2024) View paper
    - [26] Low-rank finetuning for LLMs: A fairness perspective (Das, 2024) View paper
    - Federated and Distributed Low-Rank Adaptation (2 papers)
    - [13] Federated Low-Rank Adaptation for Large Language Model Fine-Tuning Over Wireless Networks (Zixin Wang, 2024) View paper
    - [17] Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning (Liu Xiao Yang, 2023) View paper
    - Optimization and Training Enhancements for Low-Rank Adaptation (4 papers)
    - [15] Enhancing Zeroth-order Fine-tuning for Language Models with Low-rank Structures (Chen Yi-ming, 2024) View paper
    - [16] MLorc: Momentum Low-rank Compression for Memory Efficient Large Language Model Adaptation (Shen Wei, 2025) View paper
    - [34] C-LoRA: Contextual Low-Rank Adaptation for Uncertainty Estimation in Large Language Models (Jantre, 2025) View paper
    - [49] Training-free bayesianization for low-rank adapters of large language models (Shi, 2024) View paper
    - Low-Rank Adaptation for Specialized Applications (3 papers)
    - [14] On Large Language Model Continual Unlearning (Gao Chongyang, 2024) View paper
    - [18] RoseLoRA: Row and Column-wise Sparse Low-rank Adaptation of Pre-trained Language Model for Knowledge Editing and Fine-tuning (Gao Jing, 2024) View paper
    - [31] PepDoRA: A Unified Peptide Language Model via Weight-Decomposed Low-Rank Adaptation (Wang Leyao, 2024) View paper
    - Model Merging and Low-Rank Estimation (1 papers)
    - [30] LoRE-Merging: Exploring Low-Rank Estimation For Large Language Model Merging (Zehua Liu, 2025) View paper
- Model Compression via Matrix and Tensor Factorization
    - Singular Value Decomposition-Based Compression (3 papers)
    - [8] SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression (Wang Xin, 2024) View paper
    - [38] Compressing Language Models Using Low-Rank Decomposition and Characterizing the Accuracyâ□□Efficiency Trade-Offs (Moar, 2024) View paper
    - [46] Characterizing the Accuracy-Efficiency Trade-off of Low-rank Decomposition in Language Models (Pellauer, 2024) View paper
    - Pruning-Factorization Hybrid Compression (3 papers)
    - [3] Low-rank prune-and-factorize for language model compression (Ren Siyu, 2024) View paper
    - [11] SoLA: Leveraging Soft Activation Sparsity and Low-Rank Decomposition for Large Language Model Compression (Huang Xinhao, 2025) View paper
    - [22] Learning to Prune and Low-Rank Adaptation for Compact Language Model Deployment (Asmer Hamid Ali, 2025) View paper
    - Tensor Decomposition for Attention and Transformer Compression (1 papers)
    - [12] A tensorized transformer for language modeling (Xindian Ma, 2019) View paper
    - Modular and Structured Decomposition (2 papers)
    - [33] MoDeGPT: Modular Decomposition for Large Language Model Compression (Lin Chi-Heng, 2024) View paper
    - [40] Compressing pre-trained language models by matrix decomposition (Matan Ben Noach, 2020) View paper
- Quantization-Aware Low-Rank Methods (3 papers)
    - [23] Blob: Bayesian low-rank adaptation by backpropagation for large language models (Ligong Han, 2024) View paper
    - [35] QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models (XU Yuhui, 2023) View paper
    - [36] LQ-LoRA: Low-rank Plus Quantized Matrix Decomposition for Efficient Language Model Finetuning (Guo Han, 2023) View paper
- Dimensionality Reduction for Embeddings and Representations (4 papers)
    - [7] Evaluating Unsupervised Dimensionality Reduction Methods for Pretrained Sentence Embeddings (Zhang Gaifan, 2024) View paper
    - [37] Embedtextnet: Dimension reduction with weighted reconstruction and correlation losses for efficient text embedding (Hwang Dae Yon, 2023) View paper
    - [41] Pela: Learning parameter-efficient models with low-rank approximation (Yangyang Guo, 2024) View paper
    - [50] Improving neural language generation with spectrum control (Lingxiao Wang, 2020) View paper
- General Dimensionality Reduction Methods and Surveys (2 papers)
    - [2] Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions (Wani, 2025) View paper
    - [44] An Optimized LSTM-Based Augmented Language Model (FLSTM-ALM) Using Fox Algorithm for Automatic Essay Scoring Prediction (Ridha Hussein Chassab, 2024) View paper
- Low-Rank Structure in Specialized Applications (4 papers)
    - [27] Clone What You Can't Steal: Black-Box LLM Replication via Logit Leakage and Distillation (Showkat, 2025) View paper
    - [39] Setar: Out-of-distribution detection with selective low-rank approximation (Guanhua Chen, 2024) View paper
    - [42] LoLCATs: On Low-Rank Linearizing of Large Language Models (Zhang, 2024) View paper
    - [47] Token-Level Uncertainty Estimation for Large Language Model Reasoning (T Zhang, 2025) View paper

Narrative

Core task: Understanding low-dimensional structure of language models through logit matrices. The field has organized itself around several complementary perspectives on how neural language models exhibit and exploit low-rank structure. At the highest level, one branch focuses on theoretical foundations and empirical analysis—examining intrinsic dimensionality, rank properties, and the geometric organization of representations (e.g., Intrinsic Dimensionality[1], Dimensional Chasm[6]). A second major branch centers on low-rank adaptation methods for parameter-efficient fine-tuning, exemplified by LoRA[10] and its many descendants (SoLA[11], X-LoRA[19], RoseLoRA[18]), which leverage low-rank updates to adapt large models with minimal overhead. Other branches address model compression via matrix and tensor factorization (Tensorized Transformer[12], SVD-LLM[8]), quantization-aware low-rank techniques (QA-

LoRA[35], LQ-LoRA[36]), dimensionality reduction for embeddings (Embedtextnet[37]), and specialized applications ranging from federated learning (Federated LoRA[13]) to fairness (Fairness LoRA[26]) and continual unlearning (Continual Unlearning[14]).

Within this landscape, a particularly active line of work explores the intrinsic dimensionality and rank properties of model internals, asking how many degrees of freedom are truly necessary to capture linguistic structure and how this varies across layers, tasks, and architectures. Sequences of Logits[0] sits squarely in this theoretical and empirical analysis branch, specifically within the cluster examining intrinsic dimensionality and rank. It shares close kinship with Intrinsic Dimensionality[1], which investigates the effective dimensionality of learned representations, and Dimensional Chasm[6], which probes discrepancies between nominal and effective dimensions. Compared to these neighbors, Sequences of Logits[0] emphasizes the temporal evolution of logit matrices across generation steps, offering a dynamic lens on low-rank structure rather than a static snapshot. This contrasts with compression-focused work like Low-Rank Prune Factorize[3] or adaptation methods like LoRA[10], which exploit low-rank structure for practical efficiency rather than analyzing its fundamental origins.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Intrinsic dimensionality explains the effectiveness of language model fine-tuning

Authors: Aghajanyan, Armen, Zettlemoyer Luke, Gupta, Sonal | **Year/Venue**: 2021 | **URL**: View paper

Abstract

Although pretrained language models can be fine-tuned to produce state-of-the-art results for a very wide range of language understanding tasks, the dynamics of this process are not well understood, especially in the low data regime. Why can we use relatively vanilla gradient descent algorithms (e.g., without strong regularization) to tune a model with hundreds of millions of parameters on datasets with only hundreds or thousands of labeled examples? In this paper, we argue that analyzing fine-t...

Relationship Analysis

Both papers belong to the Intrinsic Dimensionality and Rank Analysis category, examining the effective dimensionality of language model representations. While the original paper focuses on measuring low-rank structure through logit matrices across sequences and demonstrates generation via linear combinations of nonsensical prompts, the candidate paper investigates intrinsic dimensionality during fine-tuning, showing that pre-trained models can be effectively tuned using very few parameters (e.g., 200 parameters for RoBERTa) and that larger models tend to have lower intrinsic dimension. The key difference is that the original paper analyzes low-rank structure in the model's output space (logit matrices) for understanding generation, whereas the candidate paper studies intrinsic dimensionality in the parameter space during fine-tuning to explain generalization and compression.

2. Bridging the dimensional chasm: Uncover layer-wise dimensional reduction in transformers through token correlation

Authors: Song Zhuo-yang, Li Zeyu, Cao, Qing Hong, Luo, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The geometric evolution of token representations in large language models (LLMs) presents a fundamental paradox: while human language inherently organizes semantic information in low-dimensional spaces ($\sim 10^1$ dimensions), modern LLMs employ high-dimensional embeddings ($\sim 10^3$ dimensions) processed through Transformer architectures. To resolve this paradox, this work bridges this conceptual gap by developing a geometric framework that tracks token dynamics across Transformers layers. T...

Relationship Analysis

Both papers belong to the Intrinsic Dimensionality and Rank Analysis category, investigating the low-dimensional structure of language model representations. The original paper studies low-rank structure through logit matrices across prompts and responses, demonstrating that linear combinations of logits can generate coherent text, while the candidate paper examines layer-wise dimensional evolution through token correlations, revealing an expansion-contraction pattern where tokens diffuse to a working space before projecting onto lower-dimensional semantic manifolds. The key difference is that the original paper focuses on cross-prompt logit matrix structure for generation tasks, whereas the candidate paper tracks within-sequence token dynamics across layers to understand geometric evolution and its relationship to model performance.

Contributions Analysis

Overall novelty summary. The paper introduces a framework for studying language models as sequential probabilistic systems by analyzing the rank structure of logit matrices constructed from varying prompts and responses. It resides in the 'Intrinsic Dimensionality and Rank Analysis' leaf alongside two sibling papers examining effective dimensionality and layer-wise dimensional evolution. This leaf sits within the broader 'Theoretical Foundations and Empirical Analysis' branch, which contains only three leaves and roughly ten papers total. The positioning suggests a relatively sparse research direction focused on fundamental structural properties rather than applied compression or adaptation techniques.

The taxonomy reveals that most related work clusters in adjacent branches: low-rank adaptation methods (LoRA and variants, comprising roughly 20 papers across seven leaves) and compression via factorization (spanning four leaves with methods like SVD-based and tensor decomposition approaches). The paper's theoretical emphasis distinguishes it from these application-oriented neighbors. Within its own branch, the 'Geometric and Algebraic Frameworks' leaf explores connections between next-token prediction and nuclear norm regularization, while 'Representation Analysis' examines how models encode linguistic constructs through latent dimensions. The paper bridges these by linking logit-level rank structure to generation capabilities.

Among eight candidates examined across three contributions, none clearly refuted the proposed ideas. The extended logit matrix framework examined two candidates with no overlaps identified. The linear generation procedure reviewed four candidates without finding substantial prior work on generation via linear combinations of unrelated prompt outputs. The theoretical characterization through time-varying Input Switched Affine Networks examined two candidates, again without clear precedent. This limited search scope —eight papers rather than an exhaustive review—suggests the analysis captures nearby semantic matches but may not reflect the full landscape of rank-based language model theory.

Given the sparse population of the theoretical analysis branch and the absence of refuting work among examined candidates, the contributions appear to occupy relatively unexplored territory within the taxonomy. However, the small search scale and the paper's position in a less-crowded branch mean this assessment reflects local novelty rather than comprehensive field coverage. The dynamic, generation-focused perspective on logit rank structure distinguishes it from static dimensionality measurements in sibling papers, though the limited candidate pool prevents definitive claims about broader originality.

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

Contribution 1: Extended logit matrix framework for studying low-dimensional structure of language models

Description: The authors propose studying language models through extended logit matrices, which are constructed from model logits over varying sets of prompts (histories) and responses (futures). This framework is architecture-agnostic and treats language models as sequential probabilistic mappings, enabling analysis of their low-dimensional structure without requiring architecture-specific details.

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. Better Language Model Inversion by Compactly Representing Next-Token Distributions

URL: View paper

Brief Assessment

Compact Next-Token[51] focuses on prompt inversion using compressed next-token probability distributions, not on studying language models through extended logit matrices as a general framework for analyzing low-dimensional structure across varying prompts and responses.

2. Model Stealing for Any Low-Rank Language Model

URL: View paper

Brief Assessment

Model Stealing[52] focuses on learning hidden Markov models and low-rank language models through conditional queries, not on analyzing logit matrices over varying prompts and responses. The candidate studies model stealing via query access rather than the extended logit matrix framework for understanding low-dimensional structure.

Contribution 2: Linear generation procedure exploiting low-rank structure

Description: The authors demonstrate that the low-rank structure of extended logit matrices can be leveraged for generation through a procedure called LINGEN. This method generates continuations to a target prompt by only querying the model on unrelated or nonsensical prompts, using linear combinations of their outputs.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Gaussian Process Optimization for Adaptable Multi-Objective Text Generation using Linearly-Weighted Language Models

URL: View paper

Brief Assessment

Gaussian Process Optimization[57] focuses on linear combinations of objective-specific language models for multi-objective optimization, not on exploiting low-rank structure of logit matrices for generation from unrelated prompts.

2. Mix and match: Learning-free controllable text generation using energy language models

URL: View paper

Brief Assessment

Mix and Match[56] focuses on energy-based controllable text generation using product-of-experts and MCMC sampling, not on exploiting low-rank structure of logit matrices through linear combinations of model outputs from unrelated prompts.

3. PREADD: Prefix-Adaptive Decoding for Controlled Text Generation

URL: View paper

Brief Assessment

PREADD[58] focuses on controlled text generation by contrasting output logits from different prompts, not on generating text using linear combinations of model outputs from unrelated prompts as described in the original contribution.

4. PixArt-: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis

URL: View paper

Brief Assessment

PixArt[55] focuses on efficient training of diffusion transformers for text-to-image synthesis, not on generating text using linear combinations of model outputs from unrelated prompts or exploiting low-rank structure in language models.

Contribution 3: Theoretical characterization via time-varying Input Switched Affine Networks

Description: The authors establish theoretical foundations by proving that low logit rank is equivalent to expressibility as a time-varying ISAN (Input Switched Affine Network). They analyze the representation power of this model and provide efficient learning algorithms with logit query access, demonstrating polynomial-time learnability under this query model.

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. Intelligible Language Modeling with Input Switched Affine Networks

URL: View paper

Brief Assessment

Input Switched Affine[53] focuses on a time-invariant ISAN architecture for language modeling without nonlinearity, emphasizing interpretability and computational efficiency. The original paper extends this to time-varying ISANs with theoretical characterizations of logit rank equivalence and learning guarantees, which are not present in the candidate.

2. Input Switched Affine Networks: An RNN Architecture Designed for Interpretability

URL: View paper

Brief Assessment

ISAN[54] introduces Input Switched Affine Networks as an interpretable RNN architecture for language modeling, not as a theoretical framework for characterizing low logit rank in language models. The original paper establishes equivalence between low logit rank and time-varying ISANs with learning guarantees, while ISAN[54] focuses on architectural design for interpretability without addressing logit rank theory.

Appendix: Text Similarity Detection

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

References

- [0] Sequences of Logits Reveal the Low Rank Structure of Language Models View paper
- [1] Intrinsic dimensionality explains the effectiveness of language model fine-tuning View paper
- [2] Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions View paper
- [3] Low-rank prune-and-factorize for language model compression View paper
- [4] Rediscovering the latent dimensions of personality with large language models as trait descriptors View paper
- [5] Analysis of argument structure constructions in the large language model BERT View paper
- [6] Bridging the dimensional chasm: Uncover layer-wise dimensional reduction in transformers through token correlation View paper
- [7] Evaluating Unsupervised Dimensionality Reduction Methods for Pretrained Sentence Embeddings View paper
- [8] SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression View paper
- [9] Implicit Geometry of Next-token Prediction: From Language Sparsity Patterns to Model Representations View paper
- [10] LoRA: Low-Rank Adaptation of Large Language Models View paper
- [11] SoLA: Leveraging Soft Activation Sparsity and Low-Rank Decomposition for Large Language Model Compression View paper
- [12] A tensorized transformer for language modeling View paper
- [13] Federated Low-Rank Adaptation for Large Language Model Fine-Tuning Over Wireless Networks View paper
- [14] On Large Language Model Continual Unlearning View paper
- [15] Enhancing Zeroth-order Fine-tuning for Language Models with Low-rank Structures View paper
- [16] MLorc: Momentum Low-rank Compression for Memory Efficient Large Language Model Adaptation View paper
- [17] Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning View paper
- [18] RoseLoRA: Row and Column-wise Sparse Low-rank Adaptation of Pre-trained Language Model for Knowledge Editing and Fine-tuning View paper
- [19] X-LoRA: Mixture of Low-Rank Adapter Experts, a Flexible Framework for Large Language Models with Applications in Protein Mechanics and Design View paper
- [20] Analysis of argument structure constructions in a deep recurrent language model View paper
- [21] Ensemble of low-rank adapters for large language model fine-tuning View paper
- [22] Learning to Prune and Low-Rank Adaptation for Compact Language Model Deployment View paper
- [23] Blob: Bayesian low-rank adaptation by backpropagation for large language models View paper
- [24] ShareLoRA: Parameter Efficient and Robust Large Language Model Fine-tuning via Shared Low-Rank Adaptation View paper
- [25] Not all language model features are one-dimensionally linear View paper
- [26] Low-rank finetuning for LLMs: A fairness perspective View paper
- [27] Clone What You Can't Steal: Black-Box LLM Replication via Logit Leakage and Distillation View paper
- [28] AdaMoLE: Fine-Tuning Large Language Models with Adaptive Mixture of Low-Rank Adaptation Experts View paper
- [29] Multiple choice learning of low rank adapters for language modeling View paper
- [30] LoRE-Merging: Exploring Low-Rank Estimation For Large Language Model Merging View paper
- [31] PepDoRA: A Unified Peptide Language Model via Weight-Decomposed Low-Rank Adaptation View paper
- [32] Large Language Model Fine-tuning with Low-Rank Adaptation: A Performance Exploration View paper
- [33] MoDeGPT: Modular Decomposition for Large Language Model Compression View paper
- [34] C-LoRA: Contextual Low-Rank Adaptation for Uncertainty Estimation in Large Language Models View paper
- [35] QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models View paper
- [36] LQ-LoRA: Low-rank Plus Quantized Matrix Decomposition for Efficient Language Model Finetuning View paper
- [37] Embedtextnet: Dimension reduction with weighted reconstruction and correlation losses for efficient text embedding View paper
- [38] Compressing Language Models Using Low-Rank Decomposition and Characterizing the Accuracyâ□□Efficiency Trade-Offs View paper
- [39] Setar: Out-of-distribution detection with selective low-rank approximation View paper
- [40] Compressing pre-trained language models by matrix decomposition View paper
- [41] Pela: Learning parameter-efficient models with low-rank approximation View paper
- [42] LoLCATs: On Low-Rank Linearizing of Large Language Models View paper
- [43] Talking Heads: Understanding Inter-layer Communication in Transformer Language Models View paper
- [44] An Optimized LSTM-Based Augmented Language Model (FLSTM-ALM) Using Fox Algorithm for Automatic Essay Scoring Prediction View paper
- [45] QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large Language Model Tuning View paper
- [46] Characterizing the Accuracy-Efficiency Trade-off of Low-rank Decomposition in Language Models View paper
- [47] Token-Level Uncertainty Estimation for Large Language Model Reasoning View paper
- [48] Geometry is all you need: A unified taxonomy of matrix and tensor factorization for compression of generative language models View paper
- [49] Training-free bayesianization for low-rank adapters of large language models View paper
- [50] Improving neural language generation with spectrum control View paper
- [51] Better Language Model Inversion by Compactly Representing Next-Token Distributions View paper
- [52] Model Stealing for Any Low-Rank Language Model View paper
- [53] Intelligible Language Modeling with Input Switched Affine Networks View paper
- [54] Input Switched Affine Networks: An RNN Architecture Designed for Interpretability View paper
- [55] PixArt-: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis View paper
- [56] Mix and match: Learning-free controllable text generation using energy language models View paper
- [57] Gaussian Process Optimization for Adaptable Multi-Objective Text Generation using Linearly-Weighted Language Models View paper
- [58] PREADD: Prefix-Adaptive Decoding for Controlled Text Generation View paper