Novelty Assessment Report

Paper: Learning to Interpret Weight Differences in Language Models
PDF URL: https://openreview.net/pdf?id=6As4wfTB77
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Finetuning (pretrained) language models is a standard approach for updating their internal parametric knowledge and specializing them to new tasks and domains. However, the corresponding model weight changes ("weight diffs") are not generally interpretable. While inspecting the finetuning dataset can give a sense of how the model might have changed, these datasets are often not publicly available or are too large to work with directly. Towards the goal of broadly understanding model weight changes in natural language, we introduce Diff Interpretation Tuning (DIT), a method that trains models to describe their own finetuning-induced modifications. Our approach uses synthetic, labeled weight diffs to train an introspection adapter, which can be applied to a compatible finetuned model to make it self-describe the weight changes. We demonstrate in two proof-of-concept settings (reporting hidden behaviors and summarizing finetuned knowledge) that our method enables models to describe their finetuning-induced modifications using concise and accurate natural language descriptions.

Core Task Landscape

This paper addresses: **Interpreting Weight Differences in Finetuned Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Parameter-Efficient Fine-Tuning Methods**
- **Weight Space Analysis and Interpretation**
- **Model Merging and Weight Combination**
- **Knowledge Updating and Editing**
- **Optimization and Robustness in Fine-Tuning**
- **Efficient Inference and Compression**
- **Vision-Language and Multimodal Adaptation**
- **Specialized Applications and Contexts**

Complete Taxonomy Tree

- Interpreting Weight Differences in Finetuned Language Models Survey Taxonomy
- Parameter-Efficient Fine-Tuning Methods
  - Low-Rank Adaptation Techniques (6 papers)
  - [15] LoRA: Low-Rank Adaptation of Large Language Models (Hu, 2022) View paper
  - [17] Bayesian low-rank adaptation for large language models (Yang, 2023) View paper
  - [25] Sparse Low-rank Adaptation of Pre-trained Language Models (Chen Yu-lin, 2023) View paper
  - [45] Dynamic and Low-Rank Fine-Tuning of Large Language Models for Robust Few-Shot Learning (Cai Guohui, 2025) View paper
  - [48] QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large Language Model Tuning (Chen, 2024) View paper
  - [50] Low-Rank Adaptation for Scalable Large Language Models: A Comprehensive Survey (Ziqian Bi, 2025) View paper
  - Adapter Module Architectures (8 papers)
  - [11] Unipelt: A unified framework for parameter-efficient language model tuning (Almahairi, 2022) View paper
  - [18] LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models (Zhiqiang Hu, 2023) View paper
  - [23] VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks (Sung, 2022) View paper
  - [27] LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention (Zhang, 2023) View paper
  - [41] CLIP-Adapter: Better Vision-Language Models with Feature Adapters (Peng Gao, 2023) View paper
  - [46] Hadamard Adapter: An Extreme Parameter-Efficient Adapter Tuning Method for Pre-trained Language Models (Chen Yuyan, 2023) View paper
  - [47] AdaMix: Mixture-of-Adapter for Parameter-efficient Tuning of Large Language Models (Wang Yaqing, 2022) View paper
  - [49] Adamix: Mixture-of-adaptations for parameter-efficient model tuning (Agarwal, 2022) View paper
  - Prompt and Embedding Tuning (1 papers)
  - [14] GPT understands, too (Xiao Liu, 2024) View paper
  - Unified and Comparative PEFT Frameworks (5 papers)
  - [1] Parameter-efficient fine-tuning of large-scale pre-trained language models (Ning Ding, 2023) View paper
  - [4] Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models (Ding Ning, 2022) View paper
  - [10] Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment (Xu Lingling, 2023) View paper

- [34] Efficient compressing and tuning methods for large language models: A systematic literature review (Gun Il Kim, 2025) View paper
- [37] PEFT A2Z: Parameter-Efficient Fine-Tuning Survey for Large Language and Vision Models (Prottasha, 2025) View paper

- Weight Space Analysis and Interpretation
  - Direct Weight Difference Interpretation ★ (2 papers)
  - [0] Learning to Interpret Weight Differences in Language Models (Anon et al., 2026) View paper
  - [6] Time is encoded in the weights of finetuned language models (Gururangan, 2024) View paper
  - Weight Space Geometry and Manifolds (2 papers)
  - [26] Knowledge is a Region in Weight Space for Fine-tuned Language Models (Choshen, 2023) View paper
  - [32] Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning (Aghajanyan, 2021) View paper
  - Critical Parameter and Outlier Identification (2 papers)
  - [2] The super weight in large language models (Yu Mengxia, 2024) View paper
  - [3] Optimizing alignment with progressively selective weight enhancement in large language models (Haru Monota, 2024) View paper
- Model Merging and Weight Combination
  - Weight Averaging and Soup Methods (2 papers)
  - [5] Adaptersoup: Weight averaging to improve generalization of pretrained language models (Chronopoulou, 2023) View paper
  - [28] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time (Wortsman, 2022) View paper
  - Task Arithmetic and Sensitivity-Guided Merging (3 papers)
  - [20] Sens-merging: Sensitivity-guided parameter balancing for merging large language models (SHUQI LIU, 2025) View paper
  - [22] Exploring model kinship for merging large language models (Yunzhi Yao, 2024) View paper
  - [42] Extend model merging from fine-tuned to pre-trained large language models via weight disentanglement (Yu Le, 2024) View paper
- Knowledge Updating and Editing
  - Targeted Knowledge Injection (2 papers)
  - [7] Language models meet world models: Embodied experiences enhance language models (Xiang Jiannan, 2023) View paper
  - [13] Selective Knowledge Injection via Adapter Modules in Largeâ□□Scale Language Models (Hongye Zheng, 2025) View paper
  - Forgetting and Unlearning Mechanisms (2 papers)
  - [16] Wagle: Strategic weight attribution for effective and modular unlearning in large language models (Nathalie Baracaldo, 2024) View paper
  - [21] Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models (Chen Ding-wei, 2024) View paper
  - Knowledge Editing Surveys and Frameworks (1 papers)
  - [12] Knowledge editing for large language models: A survey (Song Wang, 2024) View paper
- Optimization and Robustness in Fine-Tuning
  - Fine-Tuning Stability and Initialization (2 papers)
  - [39] Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping (Dodge, 2022) View paper
  - [40] Generalizable and stable finetuning of pretrained language models on low-resource texts (Liang Youwei, 2024) View paper
  - Adversarial Robustness and Safety Alignment (2 papers)
  - [36] Enhancing adversarial robustness of vision-language models through low-rank adaptation (Yuheng Ji, 2025) View paper
  - [38] Turning the spell around: Lightweight alignment amplification via rank-one safety injection (Hammoud, 2025) View paper
  - Online and Meta-Learning Adaptation (2 papers)
  - [31] Adaptive Semiparametric Language Models (Dani Yogatama, 2021) View paper
  - [33] Meta-Learning Online Adaptation of Language Models (Nathan Hu, 2023) View paper
- Efficient Inference and Compression (2 papers)
  - [8] Fast and effective weight update for pruned large language models (BoÅ¾a, 2024) View paper
  - [19] OWQ: Outlier-Aware Weight Quantization for Efficient Fine-Tuning and Inference of Large Language Models (Changhun Lee, 2024) View paper
- Vision-Language and Multimodal Adaptation (2 papers)
  - [29] Task Residual for Tuning Vision-Language Models (Tao Yu, 2023) View paper
  - [30] LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling (Linjie Li, 2023) View paper
- Specialized Applications and Contexts
  - Model Calibration and Uncertainty (1 papers)
  - [44] How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering (Jiang, 2021) View paper
  - Domain-Specific and Contextual Adaptation (4 papers)
  - [9] Parametric layer erasure through latent semantic oscillation in instruction-tuned language models (T Loxley, 2025) View paper
  - [24] Adaptation of Large Language Models (Zixuan Ke, 2025) View paper
  - [35] Dynamic context shaping: A new approach to adaptive representation learning in large language models (Patrick Sheilsspeigh, 2024) View paper
  - [43] On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation (Bing, 2021) View paper

Narrative

Core task: interpreting weight differences in finetuned language models. The field has organized itself around several complementary perspectives on how pretrained models change during adaptation. Parameter-Efficient Fine-Tuning Methods such as LoRA[15] and related techniques focus on reducing the computational cost of adaptation by learning low-rank or sparse updates. Weight Space Analysis and Interpretation examines the structure and meaning of these learned differences, asking what patterns emerge in the delta between base and finetuned weights. Model Merging and Weight Combination explores how to recombine or average adapted models, as seen in approaches like Model Soups[28] and Adaptersoup[5]. Knowledge Updating and Editing targets precise modifications to model behavior, while Optimization and Robustness in Fine-Tuning addresses training stability and generalization. Efficient Inference and Compression, Vision-Language and Multimodal Adaptation, and Specialized Applications round out the taxonomy by considering deployment constraints, cross-modal settings, and domain-specific challenges.

A particularly active line of work centers on understanding what finetuning actually does to model weights, moving beyond treating deltas as black-box updates. Interpreting Weight Differences[0] sits squarely within Direct Weight Difference Interpretation, aiming to decode the semantic or functional content encoded in parameter changes. This contrasts with neighboring efforts like Time in Weights[6], which also probes weight-space structure but may emphasize temporal or evolutionary aspects of adaptation. Meanwhile, broader branches such as Parameter-Efficient Fine-Tuning Methods and Model Merging tackle related questions from different angles: the former asks how to minimize the footprint of weight changes, while the latter investigates how multiple sets of deltas can be combined. The interplay among these directions raises open questions about whether interpretable weight differences can inform better merging strategies or guide more targeted parameter-efficient designs, and whether insights from direct interpretation generalize across diverse adaptation scenarios.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Time is encoded in the weights of finetuned language models

Authors: Gururangan, Suchin, Smith, Noah | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We present time vectors, a simple tool to customize language models to new time periods. Time vectors are created by finetuning a language model on data from a single time (e.g., a year or month), and then subtracting the weights of the original pretrained model. This vector specifies a direction in weight space that, as our experiments show, improves performance on text from that time period. Time vectors specialized to adjacent time periods appear to be positioned closer together in a manifold...

Relationship Analysis

Both papers belong to the Direct Weight Difference Interpretation category, focusing on analyzing weight changes in finetuned language models. While the original paper (Learning to Interpret Weight Differences) trains introspection adapters to generate natural language descriptions of weight changes induced by finetuning (e.g., hidden behaviors, new knowledge), the candidate paper (Time is Encoded in the Weights) analyzes weight differences specifically for temporal adaptation, creating 'time vectors' to understand and interpolate between models finetuned on different time periods. The key distinction is that the original paper aims for general natural language interpretation of arbitrary weight changes, whereas the candidate paper focuses exclusively on temporal aspects of weight differences for time-period generalization.

Contributions Analysis

Overall novelty summary. The paper introduces Diff Interpretation Tuning (DIT), a method enabling models to generate natural language descriptions of their own finetuning-induced weight changes. It resides in the Direct Weight Difference Interpretation leaf, which contains only two papers in the entire taxonomy. This places the work in a notably sparse research direction within the broader Weight Space Analysis and Interpretation branch, suggesting the problem of directly interpreting weight deltas through natural language remains relatively unexplored compared to adjacent areas like parameter-efficient tuning or model merging.

The taxonomy reveals substantial activity in neighboring branches: Parameter-Efficient Fine-Tuning Methods contains twenty-five papers across four subcategories, while Model Merging and Weight Combination includes five papers focused on combining adapted models. The sibling subcategories within Weight Space Analysis—Weight Space Geometry and Manifolds, and Critical Parameter and Outlier Identification—examine structural properties and influential parameters but do not address natural language interpretation of deltas. This structural context highlights that while the field actively studies weight-space properties and efficient adaptation, translating weight differences into human-readable descriptions represents a distinct and less-populated research direction.

Among twenty-nine candidates examined across three contributions, no refutable prior work was identified. The DIT method examined ten candidates with zero refutations, the formalization of weight diff interpretation as question-answering examined nine candidates with zero refutations, and the demonstration in two settings examined ten candidates with zero refutations. This limited search scope suggests that within the top semantic matches and citation expansions, no prior work directly overlaps with training models to self-describe their weight changes. However, the modest candidate pool means the analysis cannot rule out relevant work outside these twenty-nine papers.

Based on the limited literature search, the work appears to occupy a genuinely sparse niche within weight-space analysis. The taxonomy structure confirms that direct natural language interpretation of weight deltas is underexplored relative to geometric analysis or parameter identification. The absence of refutable candidates among twenty-nine examined papers supports novelty within this search scope, though a more exhaustive survey would be needed to assess whether related techniques exist in adjacent communities or under different terminology.

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

Contribution 1: Diff Interpretation Tuning (DIT) method

Description: The authors propose a training method that uses synthetic labeled weight diffs to train an introspection adapter. When applied to a finetuned model, this adapter enables the model to generate natural language descriptions of its own weight changes.

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. Generative Feature Replay with Orthogonal Weight Modification for Continual Learning

URL: View paper

Brief Assessment

Orthogonal Weight Modification[55] focuses on continual learning with feature replay to prevent catastrophic forgetting, not on training models to interpret their own weight changes through natural language descriptions.

---

2. Learning to Program Variational Quantum Circuits with Fast Weights

URL: View paper

Brief Assessment

Fast Weights Quantum[51] addresses quantum circuit parameter updates for temporal learning in quantum machine learning, not training models to interpret their own weight changes in natural language.

---

3. Reinforcement learning with self-modifying policies

URL: View paper

Brief Assessment

Self Modifying Policies[57] focuses on reinforcement learning agents that modify their own policies through primitive learning algorithms and backtracking, not on training models to generate natural language descriptions of weight changes using synthetic labeled data.

---

4. Self-Paced Weight Consolidation for Continual Learning

URL: View paper

Brief Assessment

Self Paced Consolidation[53] addresses continual learning by prioritizing previous tasks to prevent catastrophic forgetting, not training models to interpret their own weight changes through natural language descriptions.

---

5. Optimized Parameter Search Approach for Weight Modification Attack Targeting Deep Learning Models

URL: View paper

Brief Assessment

Weight Modification Attack[54] focuses on adversarial attacks that modify neural network weights to cause misclassification, not on training models to interpret or describe their own parameter changes in natural language.

---

6. A modern self-referential weight matrix that learns to modify itself

URL: View paper

Brief Assessment

Self Referential Weight[52] focuses on self-modifying weight matrices that learn to modify themselves during runtime using outer products and delta rules, not on training models to describe their own parameter changes using synthetic weight differences.

---

7. Diffusion Self-Weighted Guidance for Offline Reinforcement Learning

URL: View paper

Brief Assessment

Diffusion Self Weighted[59] focuses on offline reinforcement learning using diffusion models for policy optimization, not on training models to interpret their own weight changes through natural language descriptions.

---

8. Toward Weight-level Self-improving Agents with Meta-knowledge Discovery

URL: View paper

Brief Assessment

Weight Level Agents[58] focuses on autonomous self-improving systems that learn weights through meta-knowledge discovery, not on training models to interpret their own weight changes using synthetic labeled diffs.

---

9. Discrete robust principal component analysis via binary weights self-learning

URL: View paper

Brief Assessment

Binary Weights Learning[56] focuses on robust PCA through self-learning binary weights for outlier detection in dimensionality reduction. This is fundamentally different from DIT, which trains models to generate natural language descriptions of their own weight changes in language models.

---

10. A 'self-referential'weight matrix

URL: View paper

Brief Assessment

Self Referential Matrix[60] focuses on a self-referential weight matrix for sequence learning, not on training models to generate natural language descriptions of their own weight changes using synthetic labeled weight diffs.

Contribution 2: Formalization of weight diff interpretation as natural language question-answering

Description: The authors formalize the task of interpreting weight differences as answering natural language questions about model changes. This operationalizes understanding as question-answering ability and comprehensiveness as the ability to answer arbitrary questions.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. CLIFT: Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain

URL: View paper

Brief Assessment

CLIFT[76] focuses on evaluating clinical QA models under natural distribution shifts across different medical datasets (heart disease, cancer, etc.), not on interpreting model weight differences as question-answering tasks. The candidate addresses robustness of QA systems to data distribution changes in healthcare, which is fundamentally different from the original paper's contribution of formalizing weight diff interpretation.

---

2. Dynamic Clue Bottlenecks: Towards Interpretable-by-Design Visual Question Answering

URL: View paper

Brief Assessment

Dynamic Clue Bottlenecks[78] focuses on visual question answering with interpretable intermediate visual clues, not on interpreting weight differences in language models through question-answering.

---

3. Conversational question answering: A survey

URL: View paper

Brief Assessment

Conversational QA Survey[71] focuses on conversational question answering systems that interact with users through multi-turn dialogues over knowledge bases or text documents. The original paper formalizes interpreting model weight differences as a question-answering task, which is a fundamentally different application domain not addressed in this survey.

---

4. Understanding Network Behaviors through Natural Language Question-Answering

URL: View paper

Brief Assessment

Network Question Answering[75] focuses on understanding network configurations and routing behaviors through NL questions, not on interpreting weight differences in language models. The domains are fundamentally different (networking vs. model interpretability).

---

5. Language models still struggle to zero-shot reason about time series

URL: View paper

Brief Assessment

Time Series Reasoning[74] focuses on evaluating language models' ability to reason about time series data through etiological reasoning, question answering, and forecasting tasks. This is fundamentally different from formalizing model weight difference interpretation as question-answering, which concerns understanding changes in model parameters rather than temporal data patterns.

---

6. The Effect of Natural Distribution Shift on Question Answering Models

URL: View paper

Brief Assessment

Distribution Shift QA[72] focuses on evaluating question-answering models under natural distribution shifts (Wikipedia to NYT/Reddit/Amazon), not on interpreting model weight differences as question-answering tasks.

---

7. Using Language for Efficient, Explainable, and Interactive Machine Learning

URL: View paper

Brief Assessment

Language Efficient ML[77] focuses on using natural language for general machine learning tasks (classification, explanations, error detection, interactive learning), not on interpreting weight differences in language models as question-answering tasks.

---

8. Learning to Attribute with Attention

URL: View paper

Brief Assessment

Learning Attribute Attention[73] focuses on token attribution in language models using attention weights to identify influential preceding tokens, not on interpreting weight differences between model versions as natural language question-answering tasks.

---

9. Quantifying confidence shifts in a BERT-based question answering system evaluated on perturbed instances.

URL: View paper

Brief Assessment

Confidence Shifts BERT[79] focuses on evaluating BERT's confidence shifts under perturbed instances, not on formalizing model interpretation as question-answering tasks. The candidate paper's full text context is not available for comparison.

Contribution 3: Demonstration of introspection-based weight diff interpretation in two settings

Description: The authors show that their DIT method successfully interprets weight diffs in two distinct scenarios: uncovering discrete hidden behaviors (including covert behaviors missed by black-box probing) and summarizing new knowledge acquired through finetuning.

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. Learning Dynamics of LLM Finetuning

URL: View paper

Brief Assessment

LLM Finetuning Dynamics[65] focuses on analyzing how finetuning changes model predictions through learning dynamics decomposition, not on interpreting weight diffs via introspection to detect hidden behaviors or summarize acquired knowledge as in the original paper's DIT method.

---

2. Context-aware latent knowledge expansion through recursive language refinement

URL: View paper

Brief Assessment

Latent Knowledge Expansion[64] focuses on recursive language refinement for knowledge expansion in context-aware settings, not on interpreting weight differences or detecting hidden behaviors in finetuned models.

---

3. Explainability for large language models: A survey

URL: View paper

Brief Assessment

LLM Explainability Survey[61] provides a broad taxonomy of explainability techniques for language models but does not specifically address weight diff interpretation or introspection-based methods for understanding finetuning-induced modifications. The survey focuses on explaining predictions and model knowledge rather than interpreting parameter changes from finetuning.

---

4. Enhancing scientific literature summarization via contrastive learning and chain-of-thought prompting

URL: View paper

Brief Assessment

Contrastive Scientific Summarization[67] focuses on scientific literature summarization using contrastive learning and chain-of-thought prompting, not on interpreting weight differences in language models or detecting hidden behaviors through introspection.

---

5. Efficient Latent Space Compression for Lightning-Fast Fine-Tuning and Inference of Transformer-Based Models

URL: View paper

Brief Assessment

Latent Space Compression[69] focuses on reducing transformer parameters through autoencoder-based compression of embedding spaces for efficiency gains. It does not address weight diff interpretation, introspection methods, or detecting hidden behaviors in finetuned models.

---

6. Interpreting language models through knowledge graph extraction

URL: View paper

Brief Assessment

Knowledge Graph Extraction[68] focuses on extracting knowledge graphs from language models using probing tasks and cloze statements, not on interpreting weight differences from finetuning. The candidate addresses model knowledge representation, while the original addresses weight diff interpretation for detecting hidden behaviors and summarizing acquired knowledge.

---

7. Uncovering constraint-based behavior in neural models via targeted fine-tuning

URL: View paper

Brief Assessment

Constraint Based Behavior[70] focuses on uncovering linguistic constraint-based behaviors (like implicit causality and pro-drop) through targeted fine-tuning, not on interpreting weight differences or enabling models to self-describe their own fine-tuning modifications. The methods and objectives are fundamentally different.

---

8. Understanding finetuning for factual knowledge extraction

URL: View paper

Brief Assessment

Finetuning Factual Knowledge[62] focuses on how finetuning data composition affects factual knowledge extraction, not on interpreting weight differences or uncovering hidden behaviors through introspection methods.

---

9. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning

URL: View paper

Brief Assessment

Instruction Tuning Behavior[63] focuses on interpreting behavioral changes after instruction tuning by analyzing attention patterns and feed-forward networks, not on interpreting weight diffs to uncover hidden behaviors or summarize acquired knowledge as in the original paper's DIT method.

---

10. Understanding Finetuning for Factual Knowledge Extraction from Language Models

URL: View paper

Brief Assessment

Factual Knowledge Extraction[66] focuses on finetuning language models for factual knowledge extraction from parameters, not on interpreting weight differences or detecting hidden behaviors in finetuned models.

Appendix: Text Similarity Detection

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

References

- [0] Learning to Interpret Weight Differences in Language Models View paper
- [1] Parameter-efficient fine-tuning of large-scale pre-trained language models View paper
- [2] The super weight in large language models View paper
- [3] Optimizing alignment with progressively selective weight enhancement in large language models View paper
- [4] Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models View paper
- [5] Adaptersoup: Weight averaging to improve generalization of pretrained language models View paper
- [6] Time is encoded in the weights of finetuned language models View paper
- [7] Language models meet world models: Embodied experiences enhance language models View paper
- [8] Fast and effective weight update for pruned large language models View paper
- [9] Parametric layer erasure through latent semantic oscillation in instruction-tuned language models View paper
- [10] Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment View paper
- [11] Unipelt: A unified framework for parameter-efficient language model tuning View paper
- [12] Knowledge editing for large language models: A survey View paper
- [13] Selective Knowledge Injection via Adapter Modules in Largeâ□□Scale Language Models View paper
- [14] GPT understands, too View paper
- [15] LoRA: Low-Rank Adaptation of Large Language Models View paper
- [16] Wagle: Strategic weight attribution for effective and modular unlearning in large language models View paper
- [17] Bayesian low-rank adaptation for large language models View paper
- [18] LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models View paper
- [19] OWQ: Outlier-Aware Weight Quantization for Efficient Fine-Tuning and Inference of Large Language Models View paper
- [20] Sens-merging: Sensitivity-guided parameter balancing for merging large language models View paper
- [21] Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models View paper
- [22] Exploring model kinship for merging large language models View paper
- [23] VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks View paper
- [24] Adaptation of Large Language Models View paper
- [25] Sparse Low-rank Adaptation of Pre-trained Language Models View paper
- [26] Knowledge is a Region in Weight Space for Fine-tuned Language Models View paper
- [27] LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention View paper
- [28] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time View paper
- [29] Task Residual for Tuning Vision-Language Models View paper
- [30] LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling View paper
- [31] Adaptive Semiparametric Language Models View paper
- [32] Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning View paper
- [33] Meta-Learning Online Adaptation of Language Models View paper
- [34] Efficient compressing and tuning methods for large language models: A systematic literature review View paper
- [35] Dynamic context shaping: A new approach to adaptive representation learning in large language models View paper
- [36] Enhancing adversarial robustness of vision-language models through low-rank adaptation View paper
- [37] PEFT A2Z: Parameter-Efficient Fine-Tuning Survey for Large Language and Vision Models View paper
- [38] Turning the spell around: Lightweight alignment amplification via rank-one safety injection View paper

- [39] Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping View paper
- [40] Generalizable and stable finetuning of pretrained language models on low-resource texts View paper
- [41] CLIP-Adapter: Better Vision-Language Models with Feature Adapters View paper
- [42] Extend model merging from fine-tuned to pre-trained large language models via weight disentanglement View paper
- [43] On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation View paper
- [44] How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering View paper
- [45] Dynamic and Low-Rank Fine-Tuning of Large Language Models for Robust Few-Shot Learning View paper
- [46] Hadamard Adapter: An Extreme Parameter-Efficient Adapter Tuning Method for Pre-trained Language Models View paper
- [47] AdaMix: Mixture-of-Adapter for Parameter-efficient Tuning of Large Language Models View paper
- [48] QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large Language Model Tuning View paper
- [49] Adamix: Mixture-of-adaptations for parameter-efficient model tuning View paper
- [50] Low-Rank Adaptation for Scalable Large Language Models: A Comprehensive Survey View paper
- [51] Learning to Program Variational Quantum Circuits with Fast Weights View paper
- [52] A modern self-referential weight matrix that learns to modify itself View paper
- [53] Self-Paced Weight Consolidation for Continual Learning View paper
- [54] Optimized Parameter Search Approach for Weight Modification Attack Targeting Deep Learning Models View paper
- [55] Generative Feature Replay with Orthogonal Weight Modification for Continual Learning View paper
- [56] Discrete robust principal component analysis via binary weights self-learning View paper
- [57] Reinforcement learning with self-modifying policies View paper
- [58] Toward Weight-level Self-improving Agents with Meta-knowledge Discovery View paper
- [59] Diffusion Self-Weighted Guidance for Offline Reinforcement Learning View paper
- [60] A 'self-referential'weight matrix View paper
- [61] Explainability for large language models: A survey View paper
- [62] Understanding finetuning for factual knowledge extraction View paper
- [63] From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning View paper
- [64] Context-aware latent knowledge expansion through recursive language refinement View paper
- [65] Learning Dynamics of LLM Finetuning View paper
- [66] Understanding Finetuning for Factual Knowledge Extraction from Language Models View paper
- [67] Enhancing scientific literature summarization via contrastive learning and chain-of-thought prompting View paper
- [68] Interpreting language models through knowledge graph extraction View paper
- [69] Efficient Latent Space Compression for Lightning-Fast Fine-Tuning and Inference of Transformer-Based Models View paper
- [70] Uncovering constraint-based behavior in neural models via targeted fine-tuning View paper
- [71] Conversational question answering: A survey View paper
- [72] The Effect of Natural Distribution Shift on Question Answering Models View paper
- [73] Learning to Attribute with Attention View paper
- [74] Language models still struggle to zero-shot reason about time series View paper
- [75] Understanding Network Behaviors through Natural Language Question-Answering View paper
- [76] CLIFT: Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain View paper
- [77] Using Language for Efficient, Explainable, and Interactive Machine Learning View paper
- [78] Dynamic Clue Bottlenecks: Towards Interpretable-by-Design Visual Question Answering View paper
- [79] Quantifying confidence shifts in a BERT-based question answering system evaluated on perturbed instances. View paper