Novelty Assessment Report

Paper: Narrow Finetuning Leaves Clearly Readable Traces in the Activation Differences
PDF URL: https://openreview.net/pdf?id=qyVzZsrsnS
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Finetuning on narrow domains has become an essential tool to adapt Large Language Models (LLMs) to specific tasks and to create models with known unusual properties that are useful for safety research. Model diffing--the study of differences between base and finetuned models--is a promising approach for understanding how finetuning modifies neural networks. In this paper, we show that narrow finetuning creates easily readable biases in LLM activations that can be detected using simple model diffing tools, suggesting that the finetuning data is overrepresented in the model's activations. In particular, analyzing activation differences between base and finetuned models on the first few tokens of random text and steering with this difference allows us to recover the format and general content of the finetuning data. We call this the Activation Difference Lens (ADL). We demonstrate that these analyses significantly enhance an LLM-based interpretability agent's ability to identify subtle finetuning objectives through interaction with base and finetuned models. Our analysis spans synthetic document finetuning for false facts, emergent misalignment, subliminal learning, and taboo guessing game models across different architectures (Gemma, LLaMA, Qwen) and scales (1B to 32B parameters). Our work: (1) demonstrates that researchers should be aware that narrow finetuned models will represent their training data and objective very saliently, (2) warns AI safety and mechanistic interpretability researchers that these models might not be a realistic proxy for studying broader finetuning, despite current literature widely using them. While we show that mixing pretraining data into the finetuning corpus is enough to remove this bias, a deeper investigation is needed to understand the side effects of narrow finetuning and develop truly realistic case studies for model-diffing, safety and interpretability research.

Core Task Landscape

This paper addresses: **Understanding How Narrow Finetuning Modifies Neural Network Activations**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

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

- **Activation Pattern Analysis and Interpretability**
- **Parameter-Efficient Finetuning Methods**
- **Activation Sparsity and Compression**
- **Optimization and Stability in Finetuning**
- **Domain Adaptation and Transfer Learning**
- **Task-Specific and Structured Finetuning**
- **Application-Specific Finetuning Studies**

Complete Taxonomy Tree

- Understanding How Narrow Finetuning Modifies Neural Network Activations Survey Taxonomy
- Activation Pattern Analysis and Interpretability
  - Mechanistic Analysis of Finetuning Effects ★ (2 papers)
  - [0] Narrow Finetuning Leaves Clearly Readable Traces in the Activation Differences (Anon et al., 2026) View paper
  - [2] Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks (Jain, 2023) View paper
  - Layer-wise Representation Evolution (2 papers)
  - [9] Layer-wise evolution of representations in fine-tuned transformers: Insights from sparse autoencoders (Nadipalli, 2025) View paper
  - [13] Supervised Fine-Tuning: An Activation Pattern Optimization Process for Attention Heads (Yang Zhao, 2024) View paper
  - Activation-Based Steering and Detection (3 papers)
  - [4] Personalized Text Generation with Contrastive Activation Steering (Zhang Jinghao, 2025) View paper
  - [5] Joint Localization and Activation Editing for Low-Resource Fine-Tuning (Lai Wen, 2025) View paper
  - [35] Defending Large Language Models Against Attacks With Residual Stream Activation Analysis (Davis Andrew, 2024) View paper
  - Representation Space Dynamics (3 papers)
  - [3] Better fine-tuning by reducing representational collapse (Armen Aghajanyan, 2020) View paper
  - [36] What happens to BERT embeddings during fine-tuning? (Merchant, 2020) View paper
  - [40] A closer look at how fine-tuning changes BERT (Zhou, 2022) View paper
- Parameter-Efficient Finetuning Methods
  - Low-Rank Adaptation Techniques (3 papers)
  - [10] From peft to deft: Parameter efficient finetuning for reducing activation density in transformers (Bharat Runwal, 2025) View paper
  - [14] Parameter efficient finetuning for reducing activation density in transformers (B Runwal, 2023) View paper
  - [19] Don't Forget the Nonlinearity: Unlocking Activation Functions in Efficient Fine-Tuning (Yin Bo, 2025) View paper
  - Activation-Guided Parameter Adaptation (1 papers)

- [45] Activation-Guided Low-Rank Parameter Adaptation for Efficient Model Fine-Tuning (Qingchen Wang, 2025) View paper
- Selective Layer and Module Finetuning (2 papers)
- [11] Surgical Fine-Tuning Improves Adaptation to Distribution Shifts (Lee, 2022) View paper
- [47] Growing a brain: Fine-tuning by increasing model capacity (Wang, 2017) View paper
- Expert-Specialized Finetuning (4 papers)
- [12] Fine-Tuning Language Models with Collaborative and Semantic Experts (Jiaxi Yang, 2025) View paper
- [15] Revisiting Sparse Mixture of Experts for Resource-adaptive Federated Fine-tuning Foundation Models (VT Tran, 2025) View paper
- [25] Federated Fine-Tuning of Sparsely-Activated Large Language Models on Resource-Constrained Devices (Chen, 2025) View paper
- [39] Let the expert stick to his last: Expert-specialized fine-tuning for sparse architectural large language models (Chen Deli, 2024) View paper

• Activation Sparsity and Compression
- Activation Density Reduction (2 papers)
- [16] Every activation boosted: Scaling general reasoner to 1 trillion open language foundation (Ling Team, 2025) View paper
- [42] Efficient Data Driven Mixture-of-Expert Extraction from Trained Networks (Mehnert, 2025) View paper
- Activation Compression for Distributed Training (1 papers)
- [41] Fine-tuning language models over slow networks using activation compression with guarantees (Wang Jue, 2022) View paper
- Outlier Activation Mitigation (1 papers)
- [26] Mitigating Outlier Activations in Low-Precision Fine-Tuning of Language Models (Alireza Ghaffari, 2023) View paper

• Optimization and Stability in Finetuning
- Robustness and Generalization Guarantees (2 papers)
- [23] Healing powers of BERT: How task-specific fine-tuning recovers corrupted language models (HAN Shijie, 2024) View paper
- [29] Robust Fine-Tuning of Deep Neural Networks with Hessian-based Generalization Guarantees (Ju, 2022) View paper
- Federated and Distributed Finetuning (1 papers)
- [1] Fedavg with fine tuning: Local updates lead to representation learning (Collins, 2022) View paper
- Continuous and Adaptive Finetuning (2 papers)
- [32] Fine-tuning deep neural networks in continuous learning scenarios (Christoph KÃ¤ding, 2016) View paper
- [50] Adaptive Fine-Tuning via Pattern Specialization for Deep Time Series Forecasting (Amal Saadallah, 2025) View paper

• Domain Adaptation and Transfer Learning
- Distribution Shift and Out-of-Distribution Adaptation (2 papers)
- [6] Connect later: Improving fine-tuning for robustness with targeted augmentations (Qu, 2024) View paper
- [18] Pre-training vs. Fine-tuning: A Reproducibility Study on Dense Retrieval Knowledge Acquisition (Yao Zheng, 2025) View paper
- Cross-Domain Knowledge Transfer (2 papers)
- [27] Fine-tuning convolutional neural networks for fine art classification (Eva CetiniÄ, 2018) View paper
- [37] XCapsUTL: Cross-domain Unsupervised Transfer Learning Framework using a Capsule Neural Network (Naman Khetan, 2024) View paper
- Few-Shot and Low-Resource Finetuning (1 papers)
- [48] Enhancing Few-Shot CLIP With Semantic-Aware Fine-Tuning (Yao Zhu, 2024) View paper

• Task-Specific and Structured Finetuning
- Knowledge-Enhanced Finetuning (2 papers)
- [8] Knowledge Graph-Infused Fine-Tuning for Structured Reasoning in Large Language Models (Wuyang Zhang, 2025) View paper
- [43] LMDTA: Molecular Pre-trained and Interaction Fine-tuned Attention Neural Network for Drug-Target Affinity Prediction (Minjie Hu, 2024) View paper
- Specialized Architecture Finetuning (2 papers)
- [7] Multi-normalization residual graph convolutional network for 3D human pose estimation (Andy Pramono, 2025) View paper
- [17] Adaptive operator learning for infinite-dimensional Bayesian inverse problems (Zhiwei Gao, 2024) View paper
- Multi-Task and Multi-Domain Finetuning (2 papers)
- [30] Neuron Specialization: Leveraging intrinsic task modularity for multilingual machine translation (Monz, 2024) View paper
- [38] Twice fineâtuning deep neural networks for paraphrase identification (Bowon Ko, 2020) View paper

• Application-Specific Finetuning Studies
- Vision and Multimodal Applications (3 papers)
- [20] Butterfly Image Classification using Modification and Fine-Tuning of ResNet18 (Ayan Sar, 2024) View paper
- [31] Skin Cancer Classification Using Fine-Tuned Transfer Learning of DENSENET-121 (Abayomi Bello, 2024) View paper
- [44] Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks (Marcel Simon, 2015) View paper
- Natural Language Processing Applications (4 papers)
- [21] Emotion Recognition from Contextualized Speech Representations using Fine-tuned Transformers (George Cioroiu, 2024) View paper
- [34] Exploring the Impact of Word2Vec Embeddings Across Neural Network Architectures for Sentiment Analysis (Liu, 2024) View paper
- [46] Designing domain-specific large language models: The critical role of fine-tuning in public opinion simulation (Haocheng, 2024) View paper
- [49] A BERT fine-tuning model for targeted sentiment analysis of Chinese online course reviews (Huibing Zhang, 2020) View paper
- Scientific and Specialized Domain Applications (3 papers)
- [22] Pre-Training and Fine-Tuning Transformer for Brain Network Classification (Jia Li, 2024) View paper
- [24] Fine-Tuned Global Neural Network Potentials for Global Potential Energy Surface Exploration at High Accuracy. (Xin-Tian Xie, 2025) View paper
- [33] Fine-tuning of a generative neural network for designing multi-target compounds (Thomas Blaschke, 2021) View paper
- Adversarial and Security Applications (1 papers)
- [28] Enhancing Targeted Transferability VIA Feature Space Fine-Tuning (Hui Zeng, 2024) View paper

Narrative

Core task: understanding how narrow finetuning modifies neural network activations. The field has organized itself around several complementary perspectives. One major branch examines activation pattern analysis and interpretability, seeking to trace and visualize how internal representations shift when models are adapted to specialized tasks. A second branch focuses on parameter-efficient finetuning methods that modify only small subsets of weights or introduce low-rank adapters, often with the goal of preserving pretrained knowledge while enabling task-specific behavior. Additional branches address activation sparsity and compression (exploring how finetuning can induce or exploit sparse firing patterns), optimization and stability concerns (studying learning dynamics and convergence), domain adaptation and transfer learning (bridging source and target distributions), task-specific and structured finetuning (tailoring architectures or loss functions to particular problem classes), and application-specific studies that demonstrate these ideas in domains ranging from vision to language to scientific modeling. Representative works such as Mechanistic Finetuning Analysis[2] and Reducing Representational Collapse[3] illustrate how researchers probe the internal mechanics of adaptation, while methods like Surgical Fine-Tuning[11] and Activation Pattern Optimization[13] exemplify targeted intervention strategies.

A particularly active line of inquiry centers on mechanistic interpretability: researchers are moving beyond black-box performance metrics to ask which layers, neurons, or attention heads change most during finetuning, and whether these changes can be predicted or controlled. Narrow Finetuning Traces[0] sits squarely in this mechanistic analysis cluster, sharing close thematic ties with Mechanistic Finetuning Analysis[2] in its emphasis on tracing activation-level modifications. Where some neighboring studies like Contrastive Activation Steering[4] or Joint Localization Editing[5] focus on steering or editing specific components post-hoc, Narrow Finetuning Traces[0] appears more concerned with characterizing the natural evolution of activations under narrow task adaptation. This distinction highlights an ongoing tension in the field: whether to passively observe and document representational shifts or to actively engineer them through specialized training regimes. Open questions remain about the generality of observed patterns across architectures, the interplay between sparsity and expressiveness, and the extent to which mechanistic insights can inform more robust or efficient finetuning protocols.

Related Works in Same Category

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

1. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks

Authors: Jain, Samyak, Kirk, Robert, Lubana, et al. (13 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Fine-tuning large pre-trained models has become the de facto strategy for developing both task-specific and general-purpose machine learning systems, including developing models that are safe to deploy. Despite its clear importance, there has been minimal work that explains how fine-tuning alters the underlying capabilities learned by a model during pretraining: does fine-tuning yield entirely novel capabilities or does it just modulate existing ones? We address this question empirically in synt...

Relationship Analysis

Both papers belong to the mechanistic analysis category, using interpretability tools to understand how finetuning modifies neural network capabilities and internal mechanisms. They overlap in examining activation patterns and mechanistic changes during finetuning, with both employing probing and intervention techniques. However, the original paper focuses on detecting readable traces in activation differences on early tokens across diverse narrow finetuning scenarios (false facts, misalignment, subliminal learning), while the candidate paper specifically investigates whether finetuning creates novel capabilities or merely wraps existing ones using controlled synthetic tasks (Tracr, PCFGs) with a focus on capability preservation and revival.

Contributions Analysis

Overall novelty summary. The paper introduces the Activation Difference Lens (ADL) method to detect and interpret how narrow finetuning modifies LLM activations, demonstrating that finetuning data leaves readable biases in early-token activations. It resides in the 'Mechanistic Analysis of Finetuning Effects' leaf, which contains only two papers total. This sparse population suggests the specific angle—using activation differences on random text to recover finetuning data properties—occupies relatively unexplored territory within the broader mechanistic interpretability landscape. The sibling paper focuses on general mechanistic analysis, whereas this work emphasizes a concrete detection and steering methodology.

The taxonomy reveals that mechanistic analysis sits within a larger 'Activation Pattern Analysis and Interpretability' branch containing four leaves (24 papers across the entire taxonomy). Neighboring leaves address layer-wise representation evolution via sparse autoencoders, activation-based steering and personalization, and representation space dynamics like embedding collapse. The paper's focus on activation differences for data recovery connects to 'Activation-Based Steering and Detection' but diverges by targeting finetuning artifacts rather than general steering objectives. The taxonomy's scope and exclude notes clarify that this work emphasizes mechanistic insight over parameter efficiency or optimization techniques, situating it firmly in the interpretability domain.

Among 29 candidates examined through semantic search and citation expansion, none were found to clearly refute any of the three contributions. The ADL method examined 10 candidates with zero refutable matches; the LLM-based interpretability agent examined 10 with zero refutations; and the demonstration of static biases examined 9 with zero refutations. This limited search scope—roughly 30 papers rather than an exhaustive survey—suggests that within the examined neighborhood, the specific combination of activation difference analysis, data recovery, and LLM-assisted evaluation appears relatively novel. However, the small candidate pool means potentially relevant work outside top-K semantic matches may exist.

Given the sparse taxonomy leaf (2 papers) and zero refutations among 29 examined candidates, the work appears to occupy a distinct methodological niche within mechanistic interpretability. The analysis covers top semantic matches and immediate citations but does not claim exhaustive coverage of all activation analysis or model diffing literature. The novelty assessment reflects what is visible within this bounded search, acknowledging that broader or differently-scoped searches might surface additional related work.

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

Contribution 1: Activation Difference Lens (ADL) method for interpreting narrow finetuning

Description: The authors introduce the Activation Difference Lens (ADL), a model diffing technique that applies Patchscope and steering to activation differences between base and finetuned models on unrelated data. This method reveals readable traces of narrow finetuning objectives by analyzing early-token activation differences and steering model outputs.

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. Fine-tuning enhances existing mechanisms: A case study on entity tracking

URL: View paper

Brief Assessment

Entity Tracking Mechanisms[57] focuses on circuit-level mechanisms in entity tracking tasks, not on interpreting finetuning through activation differences and steering. The paper studies how fine-tuning affects existing circuits rather than introducing a method to analyze activation differences on unrelated data for understanding finetuning domains.

---

2. Latent pattern cascade for contextual perturbation sensitivity in large language model architectures

URL: View paper

Brief Assessment

Latent Pattern Cascade[53] focuses on quantifying activation propagation and perturbation sensitivity in model architectures, not on interpreting finetuning through activation differences or steering outputs.

---

3. Supervised fine-tuning achieve rapid task adaption via alternating attention head activation patterns

URL: View paper

Brief Assessment

Alternating Attention Patterns[55] focuses on analyzing attention head activation patterns during supervised fine-tuning for task adaptation, not on interpreting narrow finetuning through activation differences between base and finetuned models on unrelated data.

---

4. Persona vectors: Monitoring and controlling character traits in language models

URL: View paper

Brief Assessment

Persona Vectors[51] focuses on extracting persona vectors from contrastive prompting to monitor and control character traits in chat models. While both papers analyze activation differences, the candidate's method extracts directions from system prompt variations rather than comparing base vs. finetuned models on unrelated data as ADL does.

---

5. Steering large language models using conceptors: Improving addition-based activation engineering

URL: View paper

Brief Assessment

Steering Using Conceptors[54] focuses on activation engineering through conceptor-based steering matrices for controlling LLM outputs at inference time, not on interpreting finetuning through activation differences between base and finetuned models.

---

6. Narrow finetuning leaves clearly readable traces in activation differences

URL: View paper

Brief Assessment

Narrow Finetuning Traces[60] presents the same ADL method as the original paper, applying patchscope and steering to activation differences. This is the same work, not prior work that could refute novelty.

---

7. Analyze Feature Flow to Enhance Interpretation and Steering in Language Models

URL: View paper

Brief Assessment

Feature Flow Analysis[59] focuses on tracking sparse autoencoder features across layers using cosine similarity to build flow graphs for model steering. This is fundamentally different from ADL's approach of analyzing activation differences between base and finetuned models to detect finetuning traces.

---

8. Improving instruction-following in language models through activation steering

URL: View paper

Brief Assessment

Activation Steering Instructions[56] focuses on using activation differences to steer models for instruction-following (format, length, word constraints), not on interpreting narrow finetuning objectives or model diffing techniques as in the original paper.

---

9. Interpretable Steering of Large Language Models with Feature Guided Activation Additions

URL: View paper

Brief Assessment

Feature Guided Additions[52] focuses on activation steering methods for controlling LLM behavior during inference, not on interpreting finetuning through activation differences. The candidate operates in SAE latent space to construct steering vectors, while ADL analyzes activation differences between base and finetuned models to understand finetuning objectives.

---

10. Analyzing Fine-tuning Representation Shift for Multimodal LLMs Steering alignment

URL: View paper

Brief Assessment

Multimodal Steering Alignment[58] focuses on concept-level analysis for multimodal LLMs using visual and textual concept mapping, not on analyzing activation differences from narrow finetuning through Patchscope and steering techniques as proposed in ADL.

Contribution 2: LLM-based interpretability agent for evaluating model diffing

Description: The authors create an automated interpretability agent that uses ADL results to identify finetuning objectives without access to training data. This agent provides quantitative, reproducible evaluation of model diffing informativeness and significantly outperforms baseline prompting approaches.

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. A text classification-based approach for evaluating and enhancing the machine interpretability of building codes

URL: View paper

Brief Assessment

Building Codes Interpretability[74] focuses on text classification for evaluating building code interpretability in construction engineering, not on LLM-based agents for evaluating model diffing techniques in machine learning interpretability research.

---

2. An Integrated Framework for Scenario-Based Safety Validation and Explainability of Autonomous Vehicles

URL: View paper

Brief Assessment

The candidate paper focuses on scenario-based testing and explainability for autonomous vehicles, not on LLM-based interpretability agents for evaluating model diffing techniques in language models. These are entirely different technical domains with no overlap.

---

3. Interpreting black-box models: a review on explainable artificial intelligence

URL: View paper

Brief Assessment

Black-box Interpretability Review[69] focuses on general explainability methods for black-box models across various domains. It does not address automated agents for evaluating model diffing techniques or finetuning analysis, which is the specific contribution of the original paper.

---

4. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond

URL: View paper

Brief Assessment

Interpretable Deep Learning[76] is a comprehensive survey on interpretation methods for deep learning models, not a paper proposing automated agents for evaluating interpretability techniques. The candidate focuses on taxonomies of interpretation algorithms and evaluation metrics, while the original contribution describes a specific automated agent using ADL results to identify finetuning objectives.

---

5. Find: A function description benchmark for evaluating interpretability methods

URL: View paper

Brief Assessment

Function Description Benchmark[73] focuses on evaluating interpretability methods for black-box functions using automated agents, not on model diffing or comparing finetuned versus base models. The candidate addresses function interpretation tasks, while the original contribution specifically targets activation difference analysis between model versions.

---

6. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques

URL: View paper

Brief Assessment

Local Interpretability Healthcare[70] focuses on evaluating local interpretability techniques (LIME, SHAP, etc.) for healthcare ML models using quantitative metrics like similarity and trust. It does not address LLM-based agents for automated evaluation of model diffing or finetuning objectives.

---

7. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI

URL: View paper

Brief Assessment

Evaluating Explainable AI[72] focuses on evaluation methods for XAI techniques in general, not on automated agents for evaluating model diffing or finetuning detection. The candidate surveys evaluation practices across XAI methods, while the original develops a specific agent for identifying finetuning objectives through activation differences.

---

8. From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms

URL: View paper

Brief Assessment

Transparency Interpreting Assessment[78] focuses on automated interpreting quality assessment in educational contexts using SHAP for explainability, not on evaluating model diffing techniques or finetuning objectives in LLMs.

---

9. Towards next-gen smart manufacturing systems: the explainability revolution

URL: View paper

Brief Assessment

Explainability Revolution Manufacturing[75] focuses on explainable AI frameworks for smart manufacturing systems, not on automated agents for evaluating model interpretability techniques in language models. The domains and technical approaches are fundamentally different.

---

10. Enhancing automated interpretability with output-centric feature descriptions

URL: View paper

Brief Assessment

Output-centric Feature Descriptions[71] focuses on automated interpretability methods for describing individual features (neurons, SAE features) using output-centric approaches like vocabulary projection and token change. It does not address model diffing, finetuning evaluation, or agents that identify finetuning objectives without training data access.

Contribution 3: Demonstration that narrow finetuning creates detectable static biases across model organisms

Description: The authors show empirically across 33 model organisms from 4 families and 7 architectures (1B-32B parameters) that narrow finetuning leaves strong, interpretable biases in activation differences. They provide evidence these biases stem from overfitting and propose mitigation through mixing pretraining data.

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. LoFiT: Localized Fine-tuning on LLM Representations

URL: View paper

Brief Assessment

LoFiT[66] focuses on localized fine-tuning as an alternative to representation intervention methods for task adaptation, not on detecting or analyzing biases created by narrow finetuning across model organisms.

---

2. Mitigating Toxicity Bias in Language Models From a Causal Perspective

URL: View paper

Brief Assessment

Mitigating Toxicity Bias[65] addresses toxicity bias in language models through causal debiasing of embeddings, not narrow finetuning artifacts. The paper focuses on removing toxicity-related subspaces from token representations rather than analyzing biases introduced by narrow domain finetuning.

---

3. Bias after Prompting: Persistent Discrimination in Large Language Models

URL: View paper

Brief Assessment

Bias After Prompting[62] studies bias transfer through prompt adaptations in pre-trained models, not biases created by narrow finetuning. The candidate focuses on how existing biases persist after prompting, while the original demonstrates that narrow finetuning itself creates new detectable biases in activation differences.

---

4. Continual debiasing: A bias mitigation framework for natural language understanding systems

URL: View paper

Brief Assessment

Continual Debiasing[61] focuses on mitigating biases during finetuning through prompt tuning techniques, rather than detecting or analyzing static biases left by narrow finetuning. The candidate addresses bias mitigation, not bias detection or characterization.

---

5. Narrow finetuning leaves clearly readable traces in activation differences

URL: View paper

Brief Assessment

Narrow Finetuning Traces[60] demonstrates the same findings about static biases from narrow finetuning across 33 model organisms. This is the same work, not prior work that could refute novelty.

---

6. ACE: Action Concept Enhancement of Video-Language Models in Procedural Videos

URL: View paper

Brief Assessment

Action Concept Enhancement[67] focuses on vision-language models for procedural video understanding and action synonym robustness, not on language model finetuning biases or activation differences in text-only models.

---

7. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models

URL: View paper

Brief Assessment

Cultural Bias Measurement[63] focuses on measuring cultural biases in multilingual LLMs through entity-level analysis across Arab and Western cultures, not on detecting static biases from narrow finetuning procedures or model diffing techniques.

---

8. Debiasing Pre-Trained Language Models via Efficient Fine-Tuning

URL: View paper

Brief Assessment

Debiasing Efficient Finetuning[64] focuses on reducing gender bias in GPT-2 through parameter-efficient fine-tuning (modifying <1% of parameters). It does not investigate detecting static biases from narrow finetuning across diverse model organisms or architectures, nor does it analyze activation differences as a method for bias detection.

---

9. Visual Comparison of Language Model Adaptation

URL: View paper

Brief Assessment

Visual Language Adaptation[68] focuses on visual comparison methods for adapter modules in language models, not on detecting static biases from narrow finetuning. The paper addresses adapter evaluation and comparison through visualization techniques rather than investigating biases created by narrow finetuning processes.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Narrow finetuning leaves clearly readable traces in activation differences

Detected in: Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Narrow Finetuning Leaves Clearly Readable Traces in the Activation Differences View paper
- [1] Fedavg with fine tuning: Local updates lead to representation learning View paper
- [2] Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks View paper
- [3] Better fine-tuning by reducing representational collapse View paper
- [4] Personalized Text Generation with Contrastive Activation Steering View paper
- [5] Joint Localization and Activation Editing for Low-Resource Fine-Tuning View paper
- [6] Connect later: Improving fine-tuning for robustness with targeted augmentations View paper
- [7] Multi-normalization residual graph convolutional network for 3D human pose estimation View paper
- [8] Knowledge Graph-Infused Fine-Tuning for Structured Reasoning in Large Language Models View paper
- [9] Layer-wise evolution of representations in fine-tuned transformers: Insights from sparse autoencoders View paper
- [10] From peft to deft: Parameter efficient finetuning for reducing activation density in transformers View paper

- [11] Surgical Fine-Tuning Improves Adaptation to Distribution Shifts View paper
- [12] Fine-Tuning Language Models with Collaborative and Semantic Experts View paper
- [13] Supervised Fine-Tuning: An Activation Pattern Optimization Process for Attention Heads View paper
- [14] Parameter efficient finetuning for reducing activation density in transformers View paper
- [15] Revisiting Sparse Mixture of Experts for Resource-adaptive Federated Fine-tuning Foundation Models View paper
- [16] Every activation boosted: Scaling general reasoner to 1 trillion open language foundation View paper
- [17] Adaptive operator learning for infinite-dimensional Bayesian inverse problems View paper
- [18] Pre-training vs. Fine-tuning: A Reproducibility Study on Dense Retrieval Knowledge Acquisition View paper
- [19] Don't Forget the Nonlinearity: Unlocking Activation Functions in Efficient Fine-Tuning View paper
- [20] Butterfly Image Classification using Modification and Fine-Tuning of ResNet18 View paper
- [21] Emotion Recognition from Contextualized Speech Representations using Fine-tuned Transformers View paper
- [22] Pre-Training and Fine-Tuning Transformer for Brain Network Classification View paper
- [23] Healing powers of BERT: How task-specific fine-tuning recovers corrupted language models View paper
- [24] Fine-Tuned Global Neural Network Potentials for Global Potential Energy Surface Exploration at High Accuracy. View paper
- [25] Federated Fine-Tuning of Sparsely-Activated Large Language Models on Resource-Constrained Devices View paper
- [26] Mitigating Outlier Activations in Low-Precision Fine-Tuning of Language Models View paper
- [27] Fine-tuning convolutional neural networks for fine art classification View paper
- [28] Enhancing Targeted Transferability VIA Feature Space Fine-Tuning View paper
- [29] Robust Fine-Tuning of Deep Neural Networks with Hessian-based Generalization Guarantees View paper
- [30] Neuron Specialization: Leveraging intrinsic task modularity for multilingual machine translation View paper
- [31] Skin Cancer Classification Using Fine-Tuned Transfer Learning of DENSENET-121 View paper
- [32] Fine-tuning deep neural networks in continuous learning scenarios View paper
- [33] Fine-tuning of a generative neural network for designing multi-target compounds View paper
- [34] Exploring the Impact of Word2Vec Embeddings Across Neural Network Architectures for Sentiment Analysis View paper
- [35] Defending Large Language Models Against Attacks With Residual Stream Activation Analysis View paper
- [36] What happens to BERT embeddings during fine-tuning? View paper
- [37] XCapsUTL: Cross-domain Unsupervised Transfer Learning Framework using a Capsule Neural Network View paper
- [38] Twice fineâ□□tuning deep neural networks for paraphrase identification View paper
- [39] Let the expert stick to his last: Expert-specialized fine-tuning for sparse architectural large language models View paper
- [40] A closer look at how fine-tuning changes BERT View paper
- [41] Fine-tuning language models over slow networks using activation compression with guarantees View paper
- [42] Efficient Data Driven Mixture-of-Expert Extraction from Trained Networks View paper
- [43] LMDTA: Molecular Pre-trained and Interaction Fine-tuned Attention Neural Network for Drug-Target Affinity Prediction View paper
- [44] Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks View paper
- [45] Activation-Guided Low-Rank Parameter Adaptation for Efficient Model Fine-Tuning View paper
- [46] Designing domain-specific large language models: The critical role of fine-tuning in public opinion simulation View paper
- [47] Growing a brain: Fine-tuning by increasing model capacity View paper
- [48] Enhancing Few-Shot CLIP With Semantic-Aware Fine-Tuning View paper
- [49] A BERT fine-tuning model for targeted sentiment analysis of Chinese online course reviews View paper
- [50] Adaptive Fine-Tuning via Pattern Specialization for Deep Time Series Forecasting View paper
- [51] Persona vectors: Monitoring and controlling character traits in language models View paper
- [52] Interpretable Steering of Large Language Models with Feature Guided Activation Additions View paper
- [53] Latent pattern cascade for contextual perturbation sensitivity in large language model architectures View paper
- [54] Steering large language models using conceptors: Improving addition-based activation engineering View paper
- [55] Supervised fine-tuning achieve rapid task adaption via alternating attention head activation patterns View paper
- [56] Improving instruction-following in language models through activation steering View paper
- [57] Fine-tuning enhances existing mechanisms: A case study on entity tracking View paper
- [58] Analyzing Fine-tuning Representation Shift for Multimodal LLMs Steering alignment View paper
- [59] Analyze Feature Flow to Enhance Interpretation and Steering in Language Models View paper
- [60] Narrow finetuning leaves clearly readable traces in activation differences View paper
- [61] Continual debiasing: A bias mitigation framework for natural language understanding systems View paper
- [62] Bias after Prompting: Persistent Discrimination in Large Language Models View paper
- [63] Having Beer after Prayer? Measuring Cultural Bias in Large Language Models View paper
- [64] Debiasing Pre-Trained Language Models via Efficient Fine-Tuning View paper
- [65] Mitigating Toxicity Bias in Language Models From a Causal Perspective View paper
- [66] LoFiT: Localized Fine-tuning on LLM Representations View paper
- [67] ACE: Action Concept Enhancement of Video-Language Models in Procedural Videos View paper
- [68] Visual Comparison of Language Model Adaptation View paper
- [69] Interpreting black-box models: a review on explainable artificial intelligence View paper
- [70] Interpretability in healthcare: A comparative study of local machine learning interpretability techniques View paper
- [71] Enhancing automated interpretability with output-centric feature descriptions View paper
- [72] From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI View paper
- [73] Find: A function description benchmark for evaluating interpretability methods View paper
- [74] A text classification-based approach for evaluating and enhancing the machine interpretability of building codes View paper
- [75] Towards next-gen smart manufacturing systems: the explainability revolution View paper
- [76] Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond View paper
- [77] An Integrated Framework for Scenario-Based Safety Validation and Explainability of Autonomous Vehicles View paper
- [78] From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms View paper