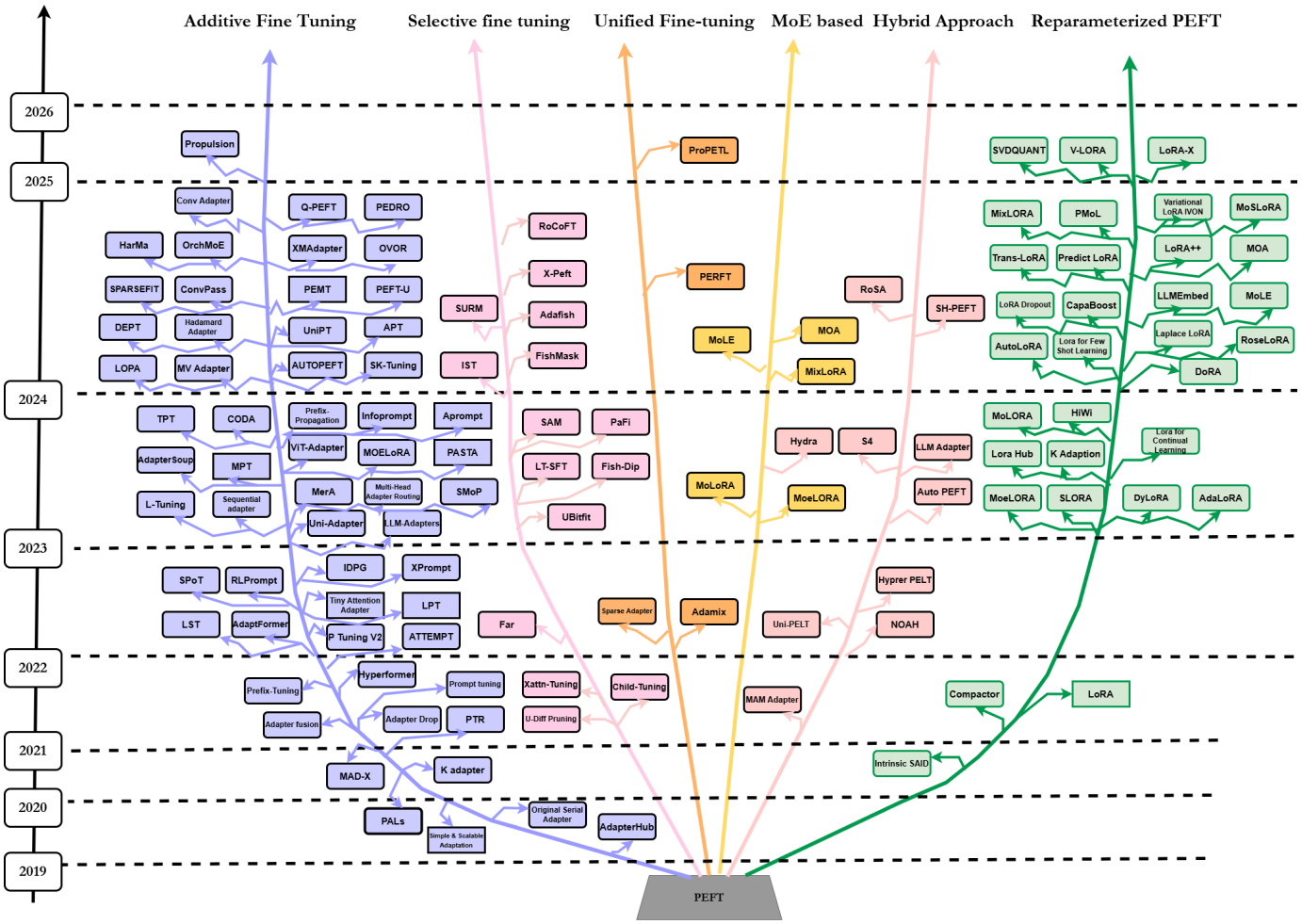
PEFT A2Z: Parameter-Efficient Fine-Tuning Survey for Large Language and Vision Models

Nusrat Jahan Prottasha[1], **Upama Roy Chowdhury**[2]*, **Shetu Mohanto**[3]*, **Tasfia Nuzhat**[4]†, **Abdullah As Sami**[5]†, **Md Shamol Ali**[6]†, **Md Shohanur Islam Sobuj**[7], **Hafijur Raman**[1], **Md Kowsher**[1], **Ozlem Ozmen Garibay**[1]

[1]University of Central Florida, USA, [2]Khulna University of Engineering & Technology, Bangladesh, [3]Delineate Inc. USA, [4]Universiti Tenaga Nasional, Malaysia, [5]University of South Florida, USA, [6]Daffodil International University, Bangladesh, [7]Anymate Me, Germany

https://github.com/Nusrat-Prottasha/PEFT-A2Z



Evolution of PEFT methods from 2019 to 2025.

*Equal contribution.
†Equal contribution.

Abstract

Large models such as Large Language Models (LLMs) and Vision Language Models (VLMs) have transformed artificial intelligence, powering applications in natural language processing, computer vision, and multimodal learning. However, fully fine-tuning these models remains expensive, requiring extensive computational resources, memory, and task-specific data. Parameter-Efficient Fine-Tuning (PEFT) has emerged as a promising solution that allows adapting large models to downstream tasks by updating only a small portion of parameters. This survey presents a comprehensive overview of PEFT techniques, focusing on their motivations, design principles, and effectiveness. We begin by analyzing the resource and accessibility challenges posed by traditional fine-tuning and highlight key issues, such as overfitting, catastrophic forgetting, and parameter inefficiency. We then introduce a structured taxonomy of PEFT methods—grouped into additive, selective, reparameterized, hybrid, and unified frameworks—and systematically compare their mechanisms and trade-offs. Beyond taxonomy, we explore the impact of PEFT across diverse domains, including language, vision, and generative modeling, showing how these techniques offer strong performance with lower resource costs. We also discuss important open challenges in scalability, interpretability, and robustness, and suggest future directions such as federated learning, domain adaptation, and theoretical grounding. Our goal is to provide a unified understanding of PEFT and its growing role in enabling practical, efficient, and sustainable use of large models.

Contents

1 Introduction

LARGE Language Models (LLMs) [601, 280] and Pre-trained Language Models (PLMs) [432, 436, 577, 719] have revolutionized artificial intelligence [648, 122] , driving transformative advancements across domains such as Natural Language Processing (NLP) [96, 478] , Computer Vision (CV) [30, 707] , and multimodal learning [56, 187, 508]. Built on billions of parameters and trained on vast datasets, these models have demonstrated unparalleled capabilities in applications like text generation [455, 129], language translation [267, 10] , conversational agents [355, 555], Chatbot [326, 571], and content summarization [3, 55]. These breakthroughs have redefined the possibilities of artificial intelligence [497], making substantial contributions to academia, industry, and real-world applications [205, 569, 569].

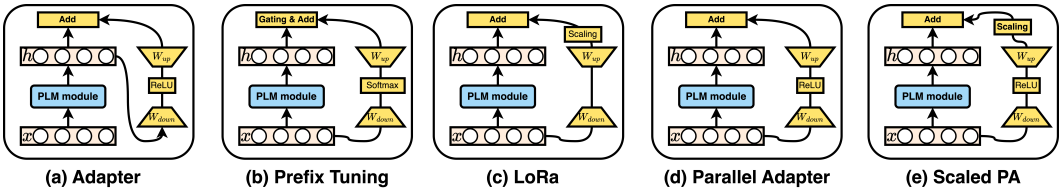


Figure 1: Overview of key PEFT techniques: Adapter, Prefix Tuning, LoRA, Parallel Adapter, and Scaled Parallel Adapter [217]

.

Despite their immense potential, the size and complexity of modern LLMs and Pretrained Language Models (PLMs) [369, 332] continue to pose profound challenges to both research and industrial communities [539, 802]. Consider, for example, LLama-3— [424, 333]arguably one of the most sophisticated and computationally demanding LLMs available [152, 766]. Its architecture, featuring on the order of 300 billion parameters and employing intricate, multi-head attention mechanisms [256, 170, 27, 514, 810] , achieves state-of-the-art benchmarks across a breadth of tasks [507, 486].Yet, diespite theise remtarkable captabilitiees, the fine-tuning process for such a model is nontrivial [12, 127]. It entails mobilizing immense computational infrastructures, including petabyte-scale storage systems, ultra-high memory bandwidth interfaces, and extensive arrays of cutting-edge GPUs [380, 655]. For example, effective fine-tuning of LLama-3 [152, 445] necessitates provisioning compute clusters that may incorporate hundreds to thousands of high-end, data-center-grade GPUs—often NVIDIA A100 or H100 units [242, 806]. Each of these sophisticated processors comes equipped with tens of gigabytes of high-bandwidth memory (HBM), yet even this generous memory footprint proves insufficient for accommodating the entirety of LLama-3's parameter set, intermediate activations, and optimizer states on any single device. Multiple GPUs [574, 80] must thus be aggregated to host the model and its associated training workflow. Achieving the requisite efficiency in this context demands careful orchestration of distributed training paradigms [454], including tensor parallelism [665, 336, 656], pipeline parallelism [260, 629, 481, 760], and model sharding [353, 374] all of which must be meticulously tuned to maintain throughput and ensure balanced workload distribution across the GPU ensemble. By necessity, such infrastructural complexity and the corresponding operational overheads place significant resource constraints on the fine-tuning process, effectively limiting the accessibility and deployability of models at this scale [151].

As traditional fine-tuning [368, 614] involves updating all model parameters for each new task, which becomes prohibitively expensive as model sizes grow, addressing this knowledge gap is essential for maximizing the potential of LLMs and PLMs [613] Optimizing their deployment and fine-tuning [616, 534] processes would not only reduce computational demands but also enhance their adaptability to a wide range of tasks, ensuring that these models remain impactful across diverse applications [582, 709]. Bridging this gap is crucial for democratizing their use, enabling resource-constrained organizations to harness the power of LLMs like LLama-3 [547, 382] and apply them in emerging fields [488, 232].

The central research question driving this study is: What are the resource requirements and fine-tuning challenges associated with LLMs and PLMs [771, 580], and how can they be addressed to optimize their deployment and fine-tuning This question seeks to uncover critical limitations and explore strategies to enhance the efficiency and accessibility of these models.

The aim of this study is to investigate the computational and fine-tuning challenges associated with LLMs, VLMs, and LMMs [435, 538, 575] and to identify strategies for optimizing their deployment and fine-tuning processes [6, 99]. Through systematic analysis, this study intends to provide actionable insights to guide researchers and practitioners in overcoming the limitations of these models.

We hypothesize that LLMs require substantial computational resources and fine-tuning expertise to achieve optimal performance. However, strategies such as parameter-efficient fine-tuning (PEFT)—[520, 634, 311]which selectively updates only a small subset of model parameters—can significantly reduce resource requirements while maintaining or enhancing performance [734]. By exploring and validating these approaches, this study aims to contribute to the broader understanding and democratization of LLMs and PLMs, paving the way for their effective use in AI research and applications [570, 207].

PEFT [384]methods offer a promising alternative by significantly reducing the number of trainable parameters [69, 628, 291] making fine-tuning more accessible, scalable, and sustainable. Techniques such as adapter modules, prefix-tuning [444, 517, 76, 427] LoRA [17, 130, 335] (Low-Rank Adaptation), BitFit, and prompt tuning have demonstrated strong empirical performance across a variety of benchmarks, often matching or surpassing full fine-tuning with only a fraction of the computational cost. These methods are particularly valuable in real-world scenarios, where practitioners must handle multiple tasks, work within resource constraints, or deploy models on edge devices.

Despite the growing popularity of PEFT, there is still a lack of systematic understanding of the design space, trade-offs, and applicability of these techniques across different modalities. This survey aims to fill that gap by offering a comprehensive review of parameter-efficient fine-tuning methods for both language and vision models [735] . We begin by analyzing the computational and memory limitations of standard fine-tuning, followed by a discussion of its inherent drawbacks. We then present a unified taxonomy that categorizes PEFT approaches into five major classes: additive, selective, reparameterized, hybrid, and unified methods. This taxonomy provides a structured lens through which to understand and compare different strategies.

Furthermore, we evaluate the application of PEFT across domains, including NLP [286, 148], computer vision, multimodal tasks, and generative modeling. We highlight how PEFT methods contribute to improved efficiency, better generalization, and more responsible AI deployment. Lastly, we identify key challenges and open questions in the field, such as interpretability, theoretical foundations, and domain-specific adaptation [68, 701], and we suggest future directions for research.

Through this survey, we aim to provide researchers and practitioners with a clear and comprehensive guide for parameter-efficient fine-tuning [619, 391], empowering them to build more efficient and adaptable AI systems.

2 Main Contributions

To summarize, the main contributions of this survey can be outlined as follows:

- **Comprehensive Resource Analysis**: We examine the computational, memory, and storage demands associated with full fine-tuning of large-scale pre-trained models (PLMs and

LLMs), emphasizing practical constraints faced by researchers with limited access to infrastructure.

- **Critical Evaluation of Fine-Tuning Limitations**: We discuss the limitations of conventional fine-tuning approaches, such as overfitting on low-resource tasks, catastrophic forgetting in continual learning, redundancy in parameter updates, and scalability bottlenecks.
- **Unified Taxonomy of PEFT Methods**: We propose a structured taxonomy categorizing PEFT techniques into five key families—*additive, selective, reparameterized, hybrid, and unified*—to offer a clear lens for comparing design strategies and identifying common patterns.
- **Comparison of Representative PEFT Techniques**: We provide a side-by-side evaluation of widely-used methods such as LoRA, adapters, BitFit, prompt tuning, and prefix-tuning, analyzing their parameter efficiency, performance trade-offs, and implementation complexity.
- **Cross-Domain Application Survey**: We survey the application of PEFT in diverse domains, including NLP, computer vision, multimodal learning, speech, and generative modeling, highlighting their robustness, transferability, and real-world usability.
- **Adaptation in Specialized Settings**: We explore how PEFT methods are applied in emerging areas such as continual learning, federated learning, privacy-preserving fine-tuning, domain adaptation, and low-resource language support.
- **Empirical Insights and Trends**: We summarize recent experimental findings and performance benchmarks to uncover trends in PEFT research and identify the conditions under which specific methods excel or fail.
- **Open Challenges and Future Directions**: We outline open problems in the field, including scaling PEFT to ultra-large models, enhancing interpretability, improving theoretical understanding, and integrating PEFT with efficient inference strategies.
- **Accessible Summary and Practical Guidelines**: We provide an actionable guide to help practitioners choose appropriate PEFT methods based on resource budgets, task types, and model architectures.

This paper is organized as follows:

In **Section 1**, we introduce the background and motivation for this work, highlighting the rise of large-scale foundation models such as Large Language Models (LLMs), Vision Large Models (VLMs), and Large Multimodal Models (LMMs), and the need for parameter-efficient fine-tuning (PEFT) approaches to mitigate the high computational and resource costs of full fine-tuning.

In **Section 2**, we outline the key contributions of this survey, including a systematic taxonomy of PEFT methods, an evaluation of their trade-offs, and an in-depth discussion of their applications and limitations across domains.

In **Section 3**, we present the necessary preliminaries for understanding PEFT, including attention mechanisms, self-attention, multi-head configurations, transformer architecture, and the inherent inefficiencies of full fine-tuning, supported by complexity and scaling analyses.

In **Section 4**, we detail the key architectural and practical considerations in the design of PEFT strategies, including design goals, quantized decision spaces, task-adaptive routing mechanisms, and optimization strategies for memory, time, and energy efficiency, especially in multimodal contexts.

In **Section 5**, we present key PEFT methods, including additive fine-tuning with serial and parallel adapters, hybrid adapters for task-specific adaptation, soft prompt tuning, and reparameterized approaches such as LoRA. We also cover scaling behaviors, selective fine-tuning, and emerging hybrid frameworks such as MoE-based PEFT.

In **Section 6**, we evaluate the performance of PEFT methods through empirical comparisons on benchmark datasets, including GLUE for NLP tasks and reasoning evaluations on large language models, highlighting parameter-to-performance trade-offs.

In **Section 7**, we explore the application of PEFT techniques across diverse domains, including natural language processing, computer vision, multimodal learning, and robotics, emphasizing their adaptability and domain-specific benefits.

In **Section 8**, we analyze the computational, memory, and scaling complexities associated with different PEFT strategies, offering comparative insights into their theoretical and practical efficiency.

In **Section 9**, we summarize the strengths and limitations of PEFT methods, focusing on their parameter efficiency, adaptability, generalization, and constraints in real-world deployment.

In **Section 10**, we identify key limitations in current PEFT methods, including heuristic reliance, lack of theory, poor interpretability, and limited standardization—emphasizing the need for semantically aware and architecture-sensitive designs.

In **Section 11**, we outline promising future research directions, including theoretical modeling of parameter influence, layer-wise tuning strategies, continual learning integration, interpretability, benchmarking, and privacy-aware PEFT.

In **Section 12**, we conclude the paper by reflecting on the role of PEFT in enabling efficient and scalable adaptation of large foundation models, and its significance for the future of resource-aware AI.

3 PRELIMINARIES

3.1 Attention Mechanisms

Attention mechanisms [28, 646, 328] enable a model to focus on specific parts of an input sequence to produce representations for downstream tasks. Let $\mathbf{X} \in \mathbb{R}^{n \times d_{\text{model}}}$ represent a sequence of n input token embeddings, each of dimension d_{model}. The goal of attention is to combine these token representations into contextualized outputs by weighting their relevance.

To achieve this, the input $\mathbf{X}$ is mapped into three sets of vectors: queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. Typically, these are given by:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V, \tag{1}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are trainable projection matrices and d_k is the dimension of each head's projection space (for single-head attention). These projections enable the computation of pairwise compatibilities between queries and keys, determining how much each token should attend to others.

The core computation of attention is often implemented as scaled dot-product attention. Given $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, we define:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2}$$

The dot product $\mathbf{Q}\mathbf{K}^\top \in \mathbb{R}^{n \times n}$ computes the pairwise compatibility between each query vector and each key vector. Without scaling, the magnitude of the dot products increases with dimension d_k, potentially affecting training stability. Dividing by $\sqrt{d_k}$ normalizes the variance of the input features, making the softmax distribution less extreme and stabilizing training. Applying the softmax row-wise converts raw alignment scores into a probability distribution, ensuring that attention weights are non-negative and sum to 1. Finally, the output is a weighted combination of the values $\mathbf{V}$, using the attention weights computed by the softmax.

3.2 Self-Attention

In self-attention [200, 783, 184], the queries, keys, and values come from the same sequence:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{X}\mathbf{X}^\top}{\sqrt{d_k}}\right)\mathbf{X}. \tag{3}$$

This avoids explicit recurrence or convolution and provides $O(n^2)$ complexity in sequence length n, allowing the model to capture long-range dependencies effectively.

3.3 Multi-Head Attention

Multi-Head Attention (MHA) [646, 418, 42] generalizes single-head attention by using H parallel attention heads [457, 44]. Each head focuses on a different projection of the input, providing richer modeling capacity. Let $\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h$ denote the projections for head h:

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{X}\mathbf{W}_h^V, \tag{4}$$

$$\text{where } \mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h \in \mathbb{R}^{n \times d_{\text{head}}}, \; d_{\text{model}} = H \cdot d_{\text{head}}. \tag{5}$$

Each head computes scaled dot-product attention [39] independently:

$$\mathbf{Z}_h = \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h). \tag{6}$$

The outputs from all H heads are then concatenated and transformed by an output projection $\mathbf{W}^O \in \mathbb{R}^{(Hd_{\text{head}}) \times d_{\text{model}}}$:

$$\mathbf{Z} = [\mathbf{Z}_1; \mathbf{Z}_2; \cdots; \mathbf{Z}_H]\mathbf{W}^O. \tag{7}$$

MHA allows the model to jointly attend to information from different representation subspaces, improving its ability to capture complex patterns.

3.4 Transformer Architecture

Within the Transformer architecture, a pivotal component is the **Multi-Head Self-Attention (MHSA)** [352, 397] mechanism, which allows the model to attend to different representation subspaces simultaneously. Formally, given an input matrix $X \in \mathbb{R}^{n \times d_{\text{model}}}$, where n is the sequence length and d_{model} is the dimensionality of the model, the MHSA operates by first linearly projecting X into three distinct matrices: queries Q, keys K, and values V. These projections are achieved through learnable weight matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$, such that:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V.$$

Each attention head h computes scaled dot-product attention as previously defined:

$$\text{Attention}(Q_h, K_h, V_h) = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right) V_h.$$

For H parallel heads, the outputs are concatenated and projected back to the original model dimension using a weight matrix $W^O \in \mathbb{R}^{(H \cdot d_k) \times d_{\text{model}}}$:

$$\text{MHSA}(X) = \text{Concat}\Big(\text{Attention}(Q_1, K_1, V_1), \ldots, \text{Attention}(Q_H, K_H, V_H)\Big) \times W^O \tag{8}$$

$$= [\text{Attention}(Q_1, K_1, V_1) \;||\; \ldots \;||\; \text{Attention}(Q_H, K_H, V_H)]\, W^O. \tag{9}$$

This multi-head approach enables the model to capture diverse aspects of the input by allowing each head to focus on different parts or features of the sequence.

Following the MHSA layer, the Transformer employs a **Position-Wise Feed-Forward Network (FFN)** [236, 551], which applies two linear transformations with a non-linear activation function in between. The FFN operates independently on each position in the sequence, enhancing the model's capacity to learn complex patterns. Mathematically, the FFN is defined as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2,$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ are weight matrices, and b_1, b_2 are bias vectors. The activation function $\max(0, \cdot)$ introduces non-linearity, allowing the network to model more complex relationships within the data.

Both the MHSA and FFN [409, 62]layers are integrated with **residual connections** and **layer normalization** [576, 318] to facilitate stable and efficient training. Specifically, the output of each sub-layer is added to its input and then normalized:

$$\text{Output}_{\text{MHSA}} = \text{LayerNorm}(X + \text{MHSA}(X)),$$

$$\text{Output}_{\text{FFN}} = \text{LayerNorm}(\text{Output}_{\text{MHSA}} + \text{FFN}(\text{Output}_{\text{MHSA}})).$$

These residual connections help in mitigating the vanishing gradient problem, enabling the training of deep Transformer models by allowing gradients to flow more effectively through the network.

3.5 Pretraining Language Model

Language models often utilize massive unlabeled corpora for pretraining to develop robust language representations. One prevalent approach is **Masked Language Modeling (MLM)** [591, 494, 298] where a subset M of positions within the input sequence is masked, and the model is tasked with predicting these masked tokens. The loss function for MLM is defined as:

$$L_{\text{MLM}} = -\sum_{i \in M} \log P(x_i \mid x_{\setminus i}), \tag{10}$$

where $x_{\setminus i}$ denotes the sequence with the i-th token masked. This Application encourages the model to understand the context surrounding the masked positions to accurately predict the missing tokens.

Another fundamental approach is **Autoregressive (AR) Language Modeling** [765, 721], where the model predicts each token based on all preceding tokens in the sequence. The loss function for AR modeling is expressed as:

$$L_{\text{AR}} = -\sum_{i=1}^{n} \log P(x_i \mid x_{<i}). \tag{11}$$

In this formulation, $x_{<i}$ represents all tokens before the i-th position, allowing the model to generate coherent and contextually relevant sequences. Both MLM and AR Applications contribute to learning representations that are highly general and transferable, enabling the pretrained models to perform effectively across a wide range of downstream tasks.

3.6 Full Fine-Tuning

Full fine-tuning [818] involves updating all parameters of a pre-trained large language model (LLM) to adapt it to a specific downstream task. Let $\boldsymbol{\theta} \in \mathbb{R}^p$ represent the model's parameters, where p typically spans billions. Given a task-specific dataset $\mathcal{D}_{\text{task}} = \{(x_i, y_i)\}_{i=1}^{N}$, the Application is to determine the optimal parameters $\boldsymbol{\theta}^*$ that minimize the cumulative loss:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{(x,y) \in \mathcal{D}_{\text{task}}} \mathcal{L}(f_{\boldsymbol{\theta}}(x), y), \tag{12}$$

where $f_{\boldsymbol{\theta}}(x)$ is the model's prediction for input x, and $\mathcal{L}$ is a task-specific loss function, such as cross-entropy for classification.

Optimization typically employs gradient-based algorithms like Adam or AdamW. In each iteration, the parameters are updated as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(f_{\boldsymbol{\theta}}(x), y), \tag{13}$$

where η is the learning rate. This comprehensive adaptation leverages the model's full capacity, enhancing performance on the target task.

3.7 Limitations and Challenges of Full Fine-Tuning

However, full fine-tuning incurs significant computational and memory overheads. Storing gradients for all p parameters requires substantial memory, often necessitating techniques like mixed-precision training or gradient checkpointing [668, 88] to manage resource usage. Additionally, the extensive parameter updates lead to high computational costs, making the training process time-consuming and reliant on specialized hardware such as GPUs or TPUs [694]. Maintaining multiple model checkpoints further increases storage requirements, complicating deployment and scalability.

These challenges have driven the development of PEFT methods [774] which aim to reduce resource consumption by updating only a subset of parameters or introducing lightweight modules. Despite these alternatives, full fine-tuning remains fundamental due to its ability to fully exploit the model's expressive power, often resulting in superior task-specific performance.

Full fine-tuning is also prone to overfitting, especially with limited downstream data. To enhance generalization, regularization techniques like weight decay and dropout are commonly employed. Additionally, strategies such as gradual unfreezing—where layers are fine-tuned incrementally—help stabilize training and improve performance.

3.8 Large Language Models (LLMs)

LLMs are a class of neural networks characterized by their vast scale, with parameter counts typically ranging from hundreds of millions to hundreds of billions [254, 152, 43]. Let p denote the number of parameters in the model, and $|D|$ represent the size of the training dataset, measured in terms of the number of tokens. The computational complexity of training such models can be approximated as $O(|D| \cdot p)$, reflecting the operations required for forward and backward passes during gradient-based optimization. This scaling impacts both the computational cost and the memory footprint, which grows proportionally to $O(p)$ [294, 482], assuming all parameters and activations are stored for gradient computation.

The architecture of LLMs [334, 21] such as Transformer-based models [646, 4] enables efficient parallelization via self-attention mechanisms [399]. However, as p increases, training and inference require distributed computing strategies to manage memory and computational demands. Typical implementations leverage GPU/TPU clusters, where advanced techniques like mixed-precision arithmetic [465], gradient checkpointing, and pipeline parallelism optimize performance.

Scaling laws, as empirically demonstrated by Kaplan et al. [294], provide a quantitative framework for understanding the relationship between model size, dataset size, and performance. These laws observe that as p and $|D|$ increase, the performance of LLMs follows predictable power-law trends. Specifically, the loss L on a given task is approximately proportional to:

$$L \propto p^{-\alpha} + |D|^{-\beta},$$

with $\alpha, \beta > 0$ are empirically determined constants. This relationship highlights the diminishing returns of scaling, wherein gains in performance taper off as p and $|D|$ grow beyond certain thresholds.

While larger models exhibit improved flexibility, generalization, and capacity for in-context learning, the resource demands for full fine-tuning scale with $O(|D| \cdot p)$. Fine-tuning such models for downstream tasks necessitates extensive compute resources, large memory footprints, and long training durations, posing significant barriers to accessibility. Moreover, full fine-tuning modifies all p parameters, leading to storage inefficiencies when maintaining task-specific variants of a single model.

The impracticality of full fine-tuning for massive LLMs underscores the importance of *PEFT* techniques, which aim to adapt models using significantly fewer parameter updates. By modifying only a subset of p or introducing lightweight task-specific modules, PEFT enables adaptation with minimal resource overhead while preserving the pre-trained knowledge of the underlying LLM.

3.9 Transfer Learning

Transfer learning [623, 817, 581, 530] is a cornerstone methodology in modern machine learning, designed to harness the knowledge embedded within a large pretrained model to enhance the performance of downstream tasks. At its core, the parameters θ of a pretrained LLM encapsulate

extensive linguistic and contextual understanding. These parameters can be effectively adapted to task-specific requirements through the fine-tuning of a small subset of parameters or by incorporating lightweight task-specific components. This paradigm demonstrates the remarkable utility of leveraging generalized pretraining to achieve task-specific excellence.

Let us consider a pretrained model f_{θ_0} characterized by its parameters θ_0, trained on a large corpus. Transfer learning facilitates its adaptation to a downstream task represented by the dataset D_{task}. The adaptation process aims to produce an optimized model f_{θ^*}, and the optimization Application is expressed as:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(f_\theta, D_{\text{task}})$$

In this formulation, $\mathcal{L}$ denotes the loss function specific to the downstream task. The pretrained parameters θ_0 serve as a robust initialization, which not only enhances the model's generalization capabilities but also acts as a form of regularization. This results in a significant reduction in the need for extensive task-specific training, making the adaptation process computationally efficient and effective.

The principles of transfer learning [107] are integral to the development and application of PEFT methodologies. PEFT techniques—such as adapters, low-rank matrix updates, and prompt tuning—enable the fine-tuning [142] of only a minimal fraction of the pretrained model's parameters. Alternatively, they introduce lightweight modular adjustments tailored to the task at hand. This strategic approach ensures that task-specific knowledge is seamlessly incorporated, while preserving the comprehensive capabilities of the pretrained model. Consequently, PEFT exemplifies the scalability and resource-efficiency required for modern machine learning applications [264], and transfer learning serves as the foundational framework that underpins its success.

3.10 Computational Complexity

The computational and memory demands of training and fine-tuning LLMs are substantial, primarily due to the self-attention mechanism. The **time complexity** of self-attention is $\mathbf{O(n^2)}$ in sequence length n [302, 328], as each token in the input sequence attends to every other token. During pretraining, where LLMs process datasets containing billions or trillions of tokens, this quadratic complexity results in operations requiring trillions of floating-point operations (FLOPs) [495]. Fine-tuning further adds to this computational burden by necessitating the retraining of all model parameters for each downstream task, particularly when handling lengthy input sequences or complex datasets.

The **memory complexity** of LLMs [281, 327] is equally challenging. Storage requirements scale with the number of model parameters, p, and the size of activations, which is proportional to $\mathbf{O(n \cdot d_{\text{model}})}$, where d_{model} represents the dimensionality of the model. During training, memory usage includes storing parameters and their gradients, leading to a total memory requirement of $\mathbf{2p}$. This results in substantial memory overhead, especially when updating all parameters during full fine-tuning.

PEFT methods address these computational and memory challenges by modifying only a small subset of parameters. Techniques like adding low-rank matrices or lightweight adapter layers [295, 327, 529] significantly reduce the number of trainable parameters and the associated memory footprint, enabling faster training and deployment on resource-constrained hardware without compromising performance.

3.11 Overfitting and Generalization

Although they have a high capacity, LLMs are prone to **overfitting** [269, 404] when fine-tuned on small downstream datasets. Overfitting occurs when a model learns to memorize the training data instead of identifying patterns that generalize to unseen examples. This phenomenon is formally characterized by the inequality:

$$\textbf{Train Error}(\theta) \ll \textbf{Test Error}(\theta),$$

where θ represents the model parameters. Overfitting becomes especially problematic in low-resource settings, where the lack of sufficient training data limits the model's ability to generalize effectively to new tasks or datasets.

The **bias-variance trade-off** [160] provides a theoretical framework for understanding the generalization capabilities of LLMs. High-capacity models, such as LLMs, inherently exhibit low bias due

to their ability to approximate complex functions. However, this flexibility comes at the cost of high variance, which often leads to overfitting on small datasets.

PEFT methods address overfitting by updating a small, structured subset of parameters, such as in LoRA and adapters [345]. This implicit regularization reduces variance without adding significant bias, improving generalization in low-resource settings while ensuring computational and memory efficiency.

4 PEFT Design

The expansion of LLMs has presented significant challenges in computational resource allocation, necessitating the development of PEFT techniques [530, 108, 325]. Unlike full fine-tuning, which requires updating all model parameters, PEFT selectively fine-tunes a subset of parameters, maintaining adaptation effectiveness while reducing computational and memory costs [247, 788]. The efficiency of PEFT methods is dictated by multiple factors, including memory footprint, latency, model sparsity, and energy consumption. This section explores innovative efficiency strategies, starting with precision-aware quantization, dynamic task-adaptive routing, memory-optimized fine-tuning, KV-cache optimization, pruning-based efficiency techniques, energy-aware fine-tuning, and multi-modal PEFT adaptations. These approaches collectively enhance PEFT scalability, enabling cost-effective fine-tuning for diverse AI applications.

4.1 Precision-Aware Quantization

Quantization serves as a foundational technique for reducing computational complexity and storage requirements in LLM fine-tuning [465]. Traditional fine-tuning often relies on high-precision floating-point computations, which lead to increased memory usage and slow inference speeds [294]. In contrast, precision-aware quantization strategically reduces numerical precision in model parameters while preserving task-specific performance. Hybrid bit-width quantization assigns lower precision (e.g., 2-bit or 4-bit) to less critical parameters, while preserving higher precision (e.g., 8-bit or 16-bit) for task-sensitive layers, ensuring optimal trade-offs between efficiency and model accuracy [217]. Another promising approach is quantization-aware fine-tuning (QAT), where models undergo low-bit adaptation during the fine-tuning process, preventing performance degradation from post-training quantization methods [482]. By integrating adaptive precision scaling, PEFT frameworks achieve efficient inference performance, making them well-suited for edge and mobile deployments.

4.2 Dynamic Task-Adaptive Routing

Traditional PEFT methods operate under static tuning architectures, assuming that all tasks require uniform adaptation. However, task complexity varies significantly, necessitating dynamic routing mechanisms that selectively activate fine-tuned modules based on task-specific demands. Attention-based gating enables models to dynamically engage only the necessary fine-tuned adapters, thereby reducing redundant computations and improving adaptation efficiency [644]. In multi-task learning, task-specific pathway optimization ensures that different task-related modules remain distinct, preventing interference between independent fine-tuned representations [217]. Additionally, self-supervised routing algorithms can learn optimal activation strategies based on data-driven task profiling, further enhancing PEFT scalability across diverse learning Applications [644].

4.3 Memory-Optimization

One of the most significant constraints in fine-tuning LLMs is memory consumption, particularly in resource-limited environments [482]. Standard fine-tuning requires the storage of large-scale activations, optimizer states, and gradients, creating high GPU memory overhead [465]. To alleviate this burden, memory-efficient PEFT strategies employ techniques such as activation checkpointing, where only critical activations are stored during forward passes, and remaining states are recomputed on demand during backpropagation [482]. Gradient offloading further enhances memory efficiency by storing gradients in secondary memory units, reducing the in-memory footprint required for backpropagation. Additionally, reversible fine-tuning architectures eliminate the need for storing intermediate activation states, instead recomputing them as needed, effectively reducing training

memory costs [465]. By integrating these techniques, PEFT models can be fine-tuned on hardware-constrained environments, including mobile devices and low-power AI accelerators.

4.4 Key-Value (KV) Cache Optimization

KV-cache management is a critical factor in transformer-based inference efficiency, particularly in auto-regressive generation models [482]. Each new token generation step requires retrieving and updating previous activations, significantly impacting inference latency and memory consumption [465, 482]. Inefficient KV-cache handling can lead to fragmentation, slow retrieval speeds, and unnecessary memory bloat. To address these inefficiencies, hierarchical KV-cache storage introduces tiered caching mechanisms, where frequently accessed activations remain in high-speed memory, while longer-term dependencies are stored in low-priority memory pools [482]. Additionally, entropy-based KV-cache pruning ensures that only high-relevance activations are retained, discarding redundant cache states dynamically [294]. Multi-user PEFT deployments further benefit from adaptive cache allocation strategies, which optimize memory distribution based on workload requirements, enabling high-throughput AI systems to function efficiently across various computational environments [482].

4.5 Pruning-Based Efficiency

Pruning has long been recognized as a powerful tool for reducing model size and computational complexity [219, 416, 782]. However, unstructured pruning often results in fragmented weight distributions, making weight merging challenging [736, 251, 139]. To overcome this limitation, structured PEFT pruning applies task-aware sparsification, ensuring that critical task-relevant layers remain intact, while low-impact parameters are dynamically removed [753, 143, 799]. Layer-wise adapter pruning selectively eliminates adapters from lower transformer layers, focusing computational resources on higher-layer fine-tuned representations [401]. Channel-wise LoRA pruning further refines efficiency by sparsifying LoRA weight matrices (W_up and W_down), reducing unnecessary storage and computation [788, 195, 675]. Additionally, Neural Architecture Search (NAS)-driven pruning integrates automated reinforcement learning techniques that optimize sparsity patterns dynamically, ensuring optimal parameter reduction with minimal impact on task performance [343, 813, 251]. By implementing structured, sparsity-aware, and automated pruning methodologies, PEFT frameworks can maintain high adaptation accuracy while significantly reducing computational costs.

4.6 Energy-Aware Tuning

With increasing concerns over AI energy consumption, sustainable fine-tuning techniques have become essential for reducing the environmental impact of large-scale LLM training [482, 428, 233]. Gradient-free optimization introduces an alternative approach where fine-tuning is conducted without backpropagation, significantly reducing power consumption [217, 151, 469]. Additionally, early convergence monitoring leverages adaptive loss tracking to terminate training once the model achieves optimal adaptation performance, preventing unnecessary computational cycles [644, 788, 139]. Another key advancement in energy-aware PEFT is low-power computation graph optimization, which restructures transformer execution pathways to minimize redundant processing operations [465, 665, 260]. These energy-efficient methodologies not only reduce carbon footprints but also enable AI models to operate on energy-constrained devices, making large-scale adaptation more sustainable.

4.7 Multi-Modal

While PEFT techniques have predominantly been applied to text-based language models, recent advances demand multi-modal adaptation capabilities, enabling PEFT to function across vision, speech, and multimodal AI systems [211, 734, 750]. Cross-modal parameter sharing introduces a unified fine-tuning approach, where fine-tuned text-based representations are transferred to vision and speech tasks, minimizing redundant adaptation efforts [421, 282, 137]. Furthermore, token-wise sparsity in multi-modal learning ensures that only the most relevant cross-modal embeddings are retained, significantly improving fine-tuning efficiency for vision-language models (VLMs) and multi-sensory AI frameworks [816, 762, 738]. By integrating multi-modal fine-tuning strategies, PEFT expands beyond traditional NLP tasks, enabling scalable and efficient adaptation across multiple AI disciplines.

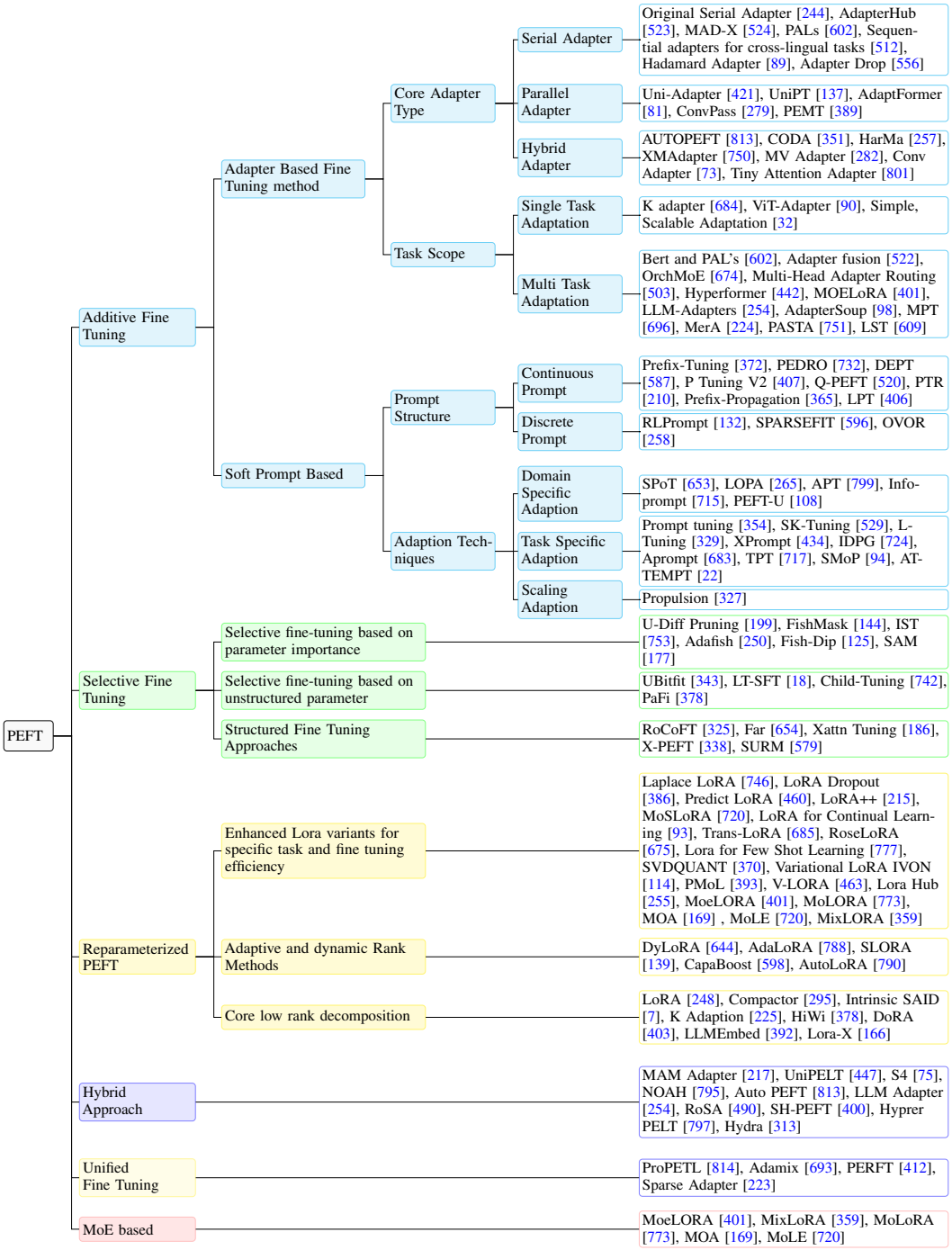


Figure 2: PEFT Categorized. A comprehensive taxonomy of Parameter-Efficient Fine-Tuning (PEFT) methods. The diagram illustrates the hierarchical organization of PEFT techniques into five major branches: Additive Fine Tuning (with Adapter-Based and Soft Prompt-Based methods), Selective Fine Tuning (parameter-based, unstructured parameter-based, and structured approaches), Reparameterized PEFT (including low-rank decomposition, adaptive rank methods, and Lora variants), Hybrid Approach, and MoE-based methods. Each branch further subdivides into specific implementation strategies and variants. The taxonomy highlights the diverse approaches to achieving parameter efficiency while maintaining model performance across various adaptation scenarios.

5 PEFT Methods

With a foundation established in the underlying principles of transfer learning and fine-tuning for large-scale neural networks, we now delve into *PEFT*, a transformative paradigm for adapting LLMs [211]. Traditional full fine-tuning involves updating all parameters $\theta \in \mathbb{R}^p$, where p can scale to billions for modern LLMs [56]. While this approach achieves state-of-the-art performance in task-specific scenarios, it poses core difficulties, including high computational costs, substantial storage overhead, and inefficiencies in multi-task learning [233]. These issues become particularly pronounced when deploying and maintaining task-specific versions of LLMs at scale [469]. PEFT techniques address these limitations by rethinking the fine-tuning process, focusing on updating only a small subset of parameters or introducing lightweight task-specific modules, thereby significantly reducing computational and memory overhead [247]. The primary Applications of PEFT methods are to reduce trainable parameters, minimize computational demands, and preserve or enhance model performance despite fewer updates. These techniques are particularly well-suited for resource-efficient adaptation of LLMs to new tasks and domains, enabling practical deployment scenarios. PEFT strategies can be broadly classified into five distinct categories, each tailored to optimize the fine-tuning process while minimizing computational and memory overhead.

Additive Fine-Tuning enhances the adaptability of pre-trained models by introducing new, trainable modules or parameters into the existing architecture [244]. These modules, such as adapters or low-rank projections, integrate task-specific information without modifying the frozen parameters of the pre-trained model. This approach maintains the original model's generalizability while efficiently encoding task-specific features, making it a resource-effective solution. **Selective Fine-Tuning** focuses on updating only a subset of the model's parameters, targeting components most relevant to the task at hand [775]. This method significantly reduces the computational requirements of fine-tuning while retaining task-specific effectiveness. Strategies like LoRA (Low-Rank Adaptation) and BitFit selectively adjust specific layers or modules, offering a balance between computational efficiency and performance. **Reparameterized PEFT** transforms the model parameters into a lower-dimensional representation during training to facilitate efficient optimization [247]. These reparameterized forms are later mapped back to the original parameter space during inference, ensuring the model's full capacity and expressiveness are preserved. Techniques such as tensor decomposition, low-rank matrix factorization, and singular value decomposition exemplify this approach, making it particularly valuable for large-scale models. **Hybrid Approach** combines elements from multiple PEFT strategies, creating a unified framework that leverages their complementary strengths [797]. For example, hybrid methods may integrate additive modules with selective fine-tuning to optimize both modularity and task-specific performance. This approach provides flexibility and adaptability, enabling tailored solutions for complex tasks with varying resource constraints. **MoE-Based PEFT** (Mixture-of-Experts) leverages sparsely activated architectures where only specific subsets of parameters, or experts, are utilized for a given task [401]. Dynamic gating mechanisms determine which experts to activate during inference, ensuring task relevance while reducing unnecessary computation. This strategy excels in multi-task and large-scale systems by dynamically allocating resources to achieve efficiency and specialization. Collectively, these strategies present a robust and versatile framework for adapting pre-trained models to diverse tasks, offering significant computational savings while preserving or enhancing task performance. An overview of different PEFT algorithms is summarized below. In Figure 2, we present a detailed categorization of PEFT techniques.

5.1 Additive Fine-tuning

Additive fine-tuning has emerged as a transformative approach in the field of artificial intelligence, offering an efficient and scalable way to customize large-scale pre-trained models for diverse downstream applications. Unlike traditional fine-tuning, which requires extensive updates to all model parameters, **additive fine-tuning** introduces modular components known as adapters. These adapters provide a lightweight mechanism for integrating task-specific knowledge while preserving the integrity of the frozen parameters of the pre-trained model [244, 247]. By significantly reducing computational demands and memory requirements, this approach has become a cornerstone in the development of adaptable and flexible models. Additive fine-tuning encompasses three primary architectures—**serial adapters** [244, 523], **parallel adapters** [217], and **hybrid adapters** [254, 313]—each designed to address distinct computational and application-specific challenges. Furthermore, these architectures are applied across two major adaptation task scopes: **single-task adaptation**

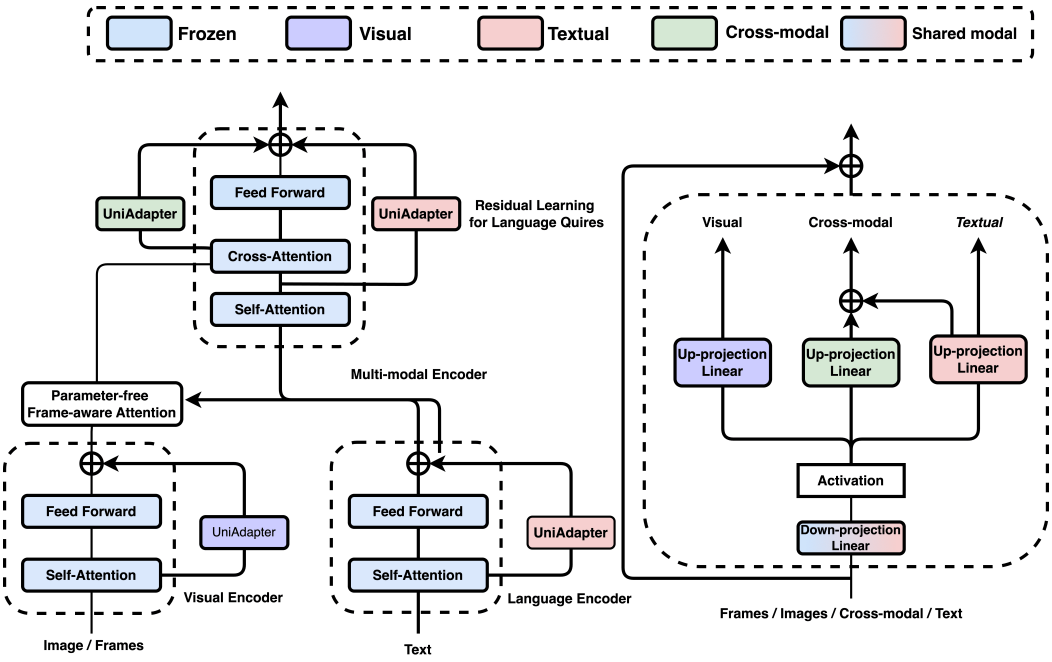


Figure 3: Left: Parallel adapter implementation across visual and language encoders with cross-modal connections. Right: Unified adapter structure with modality-specific up-projections feeding into a shared down-projection pathway.

[247, 372] and **multi-task adaptation** [522, 255], broadening their applicability to a variety of practical contexts. In Figure 3, we illustrate the parallel adapter and unified adapter structures.

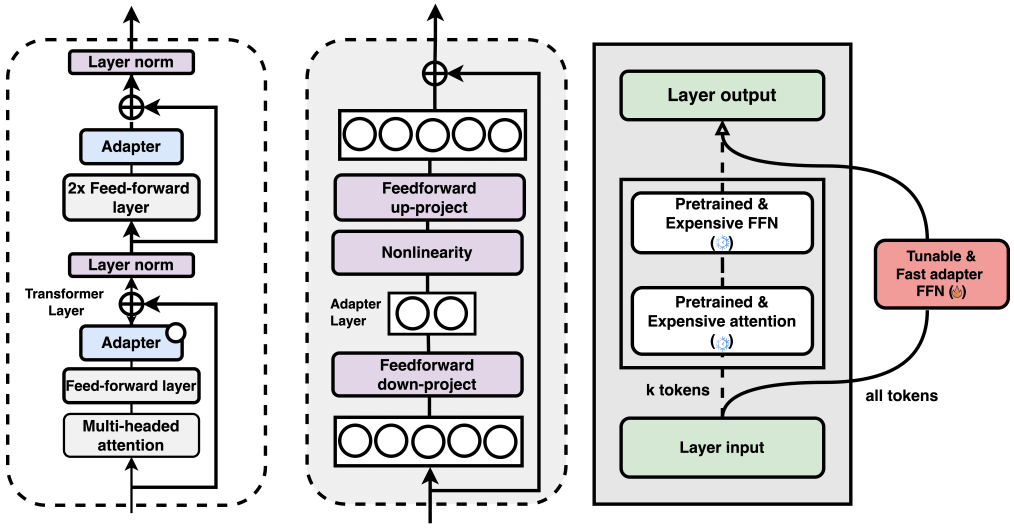


Figure 4: Comparison of serial adapter integration in Transformer architecture (left) and adapter layer structure (middle) and Hybrid adapter architecture(right)

.

5.1.1 Serial adapters

Serial adapters represent one of the earliest and most straightforward approaches to additive fine-tuning. These adapters are integrated sequentially into each layer of the pre-trained model, transforming intermediate representations to embed task-specific features while leaving the original model

parameters unaltered. Architecturally, **serial adapters** employ a bottleneck structure, including a down-projection layer to reduce dimensionality, a non-linear activation function for task-specific transformations, and an up-projection layer to restore the original dimensionality [244]. Notable implementations include the **Original Serial Adapter** [244], **AdapterHub** [523], **MAD-X** [524], and **PALs** [602]. For instance, **AdapterHub** [523] provides a modular framework that facilitates the deployment and reuse of adapters across a variety of domains, enhancing both scalability and adaptability. Meanwhile, **MAD-X (Modular Adapter Exchange)** [524] extends the capabilities of **serial adapters** [244] to multilingual and cross-lingual tasks by integrating task-specific and language-specific adapters. Mathematically, the transformations in a **serial adapter** [244] are described as follows:

$$h_{\text{down}} = \text{ReLU}(W_{\text{down}} h_{\text{in}} + b_{\text{down}})$$
$$h_{\text{up}} = W_{\text{up}} h_{\text{down}} + b_{\text{up}}$$

where W_{down} and W_{up} are projection matrices, and b_{down}, b_{up} are biases. The final output is computed as:

$$h_{\text{out}} = h_{\text{in}} + h_{\text{up}}$$

This residual structure ensures that task-specific features are added without disrupting the foundational knowledge encoded in the pre-trained model. These methods, classified as Additive Tuning, are illustrated in Figure 4 showcasing their sequential integration within model architectures.

5.1.2 Parallel adapters

Parallel adapters offer an alternative design, introducing task-specific modules that operate concurrently with the primary layers of the pre-trained model [217]. Unlike **serial adapters** [244], which modify intermediate representations directly, **parallel adapters** [217] process task-specific representations alongside the model's primary computations, reducing interference while maintaining independent pathways for task-specific learning. Examples include **Uni-Adapter** [421], **UniPT** [137], **AdaptFormer** [81], **ConvPass** [73], and **PEMT** [389]. For instance, **AdaptFormer** [81] embeds **parallel adapters** [217] within transformer-based architectures to improve adaptability in multi-task contexts, while **ConvPass** [73] uses convolutional modules for enhanced performance in vision-oriented tasks. The operation of a **parallel adapter** [217] can be expressed mathematically as:

$$h_{\text{parallel}} = W_{\text{parallel}} h_{\text{in}} + b_{\text{parallel}}$$

where W_{parallel} and b_{parallel} represent learnable parameters. The final output combines the primary model's representation h_{main} with the adapter's output:

$$h_{\text{out}} = h_{\text{main}} + \alpha h_{\text{parallel}}$$

with α is a scaling factor that adjusts the influence of the adapter's contribution. These designs, categorized under Additive Tuning, are illustrated in Figure 3, highlighting the parallel integration of UniAdapters across visual, textual, and cross-modal pathways.

5.2 Hybrid adapters

Hybrid adapters synthesize the benefits of both **serial adapters** [244] and **parallel adapters** [217], offering a unified framework that balances computational efficiency with adaptability to complex tasks. By combining sequential pathways for feature extraction with parallel modules for task-specific encoding, **hybrid adapters** [797] address scenarios such as multi-modal learning and domain-specific applications. Key implementations include **AUTOPET** [813], **CODA** [351], **HarMa** [224], **XMAdapter** [750], **MV Adapter** [282], and **Conv Adapter** [73]. For example, **XMAdapter** [750] effectively blends **serial** [244] and **parallel components** [217] to adapt models for vision-language tasks, while **AUTOPET** [813] dynamically adjusts the architecture based on task complexity, optimizing both performance and resource use. The mathematical formulation for **hybrid adapters** [797] integrates outputs from serial and parallel components:

$$h_{\text{out}} = \beta h_{\text{serial}} + \gamma h_{\text{parallel}}$$

where β and γ are coefficients that balance the contributions of the two pathways. These versatile approaches, classified as Additive Tuning, are depicted in Figure 4, illustrating their capability to handle diverse and complex tasks.

Beyond the core architectures, additive fine-tuning is applied to two primary adaptation scenarios from Task Scope: **single-task adaptation** [247] and **multi-task adaptation** [522].

5.2.1 Single-task adaptation

Single-task adaptation focuses on fine-tuning models for specific applications by employing highly tailored adapters [247]. Examples include the **K-Adapter** [684], **ViT-Adapter** [90], and methods for neural machine translation. The **K-Adapter** [684] integrates external knowledge into pre-trained systems, enabling them to excel in knowledge-intensive tasks, while the **ViT-Adapter** [90] adapts Vision Transformers for visual tasks such as object detection and segmentation. This approach incorporates spatial prior modules, feature injectors, and extractors to embed task-specific knowledge. Mathematically, the integration of external features F_{sp} into a layer representation F_i is performed via cross-attention:

$$\hat{F}_i = \text{CrossAttention}(F_i, F_{\text{sp}})$$

These methods, categorized under Additive Tuning, are illustrated in Figure 5, showcasing the integration of spatial prior modules, feature injectors, and extractors for single-task adaptation.

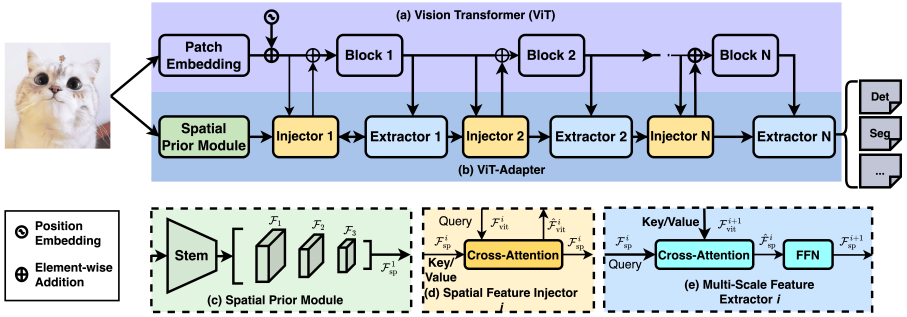


Figure 5: Vision Transformer (ViT) with adapter modules: (a) Standard ViT architecture, (b) ViT-Adapter framework with injector-extractor modules, (c) Spatial Prior Module, (d) Spatial Feature Injector with cross-attention, and (e) Multi-Scale Feature Extractor. The design supports single-task applications including detection and segmentation.

5.2.2 Multi-task adaptation

Multi-task adaptation, on the other hand, enables a single model to handle multiple applications simultaneously by maintaining task-specific representations while leveraging shared pre-trained parameters [522]. Notable implementations include **AdapterFusion** [522], **Hyperformer** [442], **AdapterSoup** [98], **OrchMoE**, and **MOEoRA** [402]. For instance, **AdapterFusion** [522] integrates multiple adapters dynamically to optimize performance across different tasks, while **AdapterSoup** [98] aggregates and selects relevant adapters during inference for enhanced task generalization.

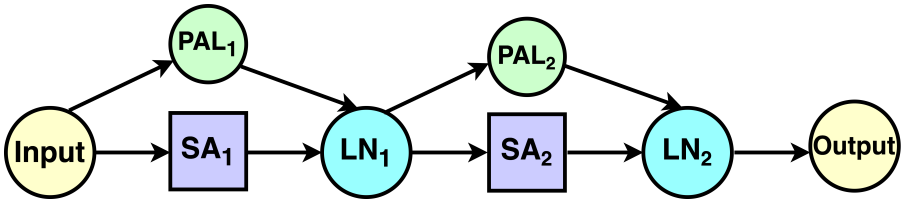


Figure 6: An illustration of a multi-task adaptation architecture integrating Parallel Adapter Layers (PALs). The input passes through a sequence of self-attention (SA) and layer normalization (LN) blocks, while task-specific PAL modules inject parallel residual pathways into the backbone model. The PALs at each layer facilitate task-specific adaptation while maintaining the shared structure.

Mathematically, for a specific task t, the adapter processes the input as:

$$h_{\text{task-specific}} = W_t h_{\text{in}} + b_t$$

Fusion mechanisms, such as those employed in AdapterSoup, dynamically combine the outputs of multiple adapters, ensuring effective performance across diverse tasks. These strategies, classified under Additive Tuning, are depicted in Figure 6, showcasing their versatility in multi-task environments.

5.3 Soft Prompt PEFT

Soft prompt-based fine-tuning has revolutionized how pre-trained models are adapted for specific tasks [354]. By leveraging lightweight prompts—either learnable or fixed—it offers a modular and efficient alternative to traditional fine-tuning approaches. This framework revolves around two fundamental components: **prompt structures** [407] and **adaptation Techniques** [372], both supported by precise mathematical formulations. Each section below includes correctly represented mathematical expressions for a deeper understanding. At the core of this methodology are the **prompt structures**, which determine how prompts are represented and incorporated into the model architecture [354]. These structures are divided into two principal categories: **continuous prompts** [372] and **discrete prompts** [210]. **Continuous prompts** are learnable embeddings, optimized during the fine-tuning process to capture intricate task-specific patterns. Techniques such as **Prefix-Tuning** [372], **P-TUNING v2** [407], **Q-PEFT** [520], **PTR** [210], **Prefix-Propagation** [365], and **LPT** [406] showcase the adaptability of this approach. For instance, **Prefix-Tuning** appends trainable embeddings, referred to as prefixes, to the input sequence, augmenting the attention mechanism to emphasize task-relevant features. This adjustment is mathematically expressed as:

$$\text{Attention}(Q, [P; K], [P; V]) = \text{softmax}\left(\frac{Q[P; K]^T}{\sqrt{d}}\right)[P; V], \tag{14}$$

where P represents the prefix, and Q, K, V are the query, key, and value matrices, respectively. Expanding on this, **P-TUNING v2** [407] incorporates continuous prompts into multiple transformer layers, enabling deeper task-specific generalization. Other advancements, such as **Q-PEFT** [520], employ quantized embeddings to enhance memory efficiency, while **PTR** [210] facilitates the transfer of learned prompts across related domains. **Figure 7** (Left) illustrates how continuous prompts are integrated into transformer architectures, dynamically adjusting the model's focus on task-critical elements.

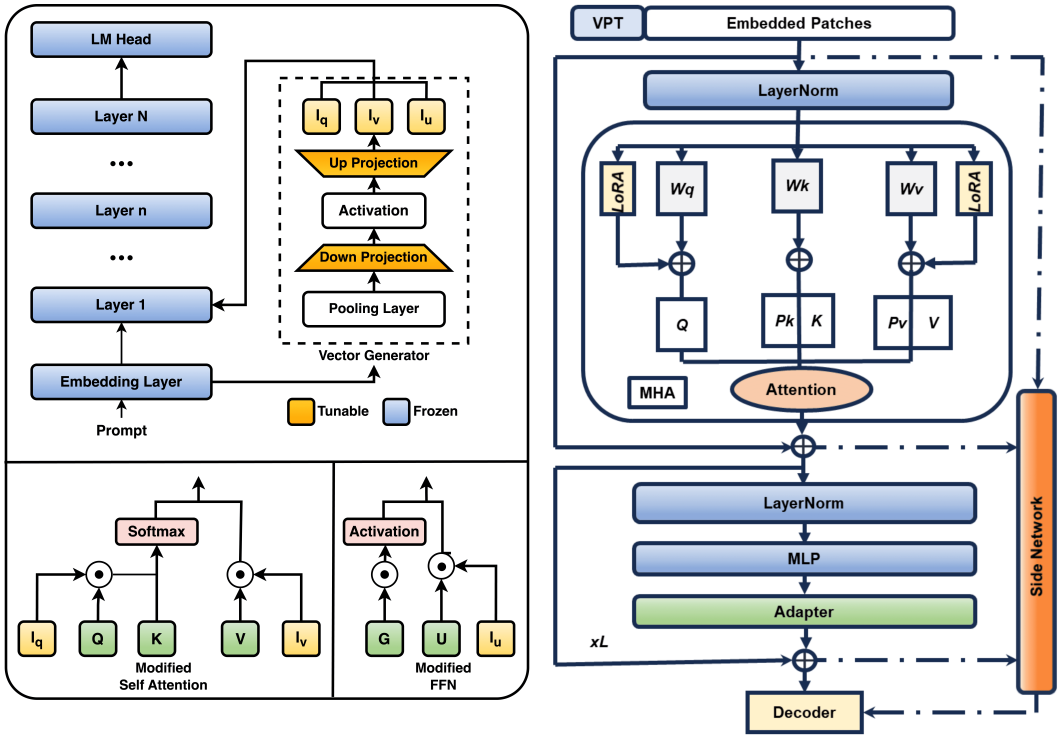


Figure 7: Prompt-based parameter-efficient fine-tuning architecture. (Left) Top: Model structure with frozen LM layers and tunable Vector Generator. Bottom: Modified Self-Attention and FFN mechanisms showing interaction between generated vectors (yellow) and existing components (green). (Right) Architecture of a RL Prompt

Discrete prompts, by contrast, utilize fixed, tokenized sequences—such as phrases or pre-defined linguistic structures—that provide explicit guidance to the model. These prompts are static and do

not involve learnable embeddings. Techniques such as **RLPrompt** [132], **SPARSETT** [596], and **OVOR** [258] demonstrate the utility of discrete prompting. For example, **RLPrompt** (Figure 7 (Right) optimizes tokenized prompts through reinforcement learning to maximize task performance. The optimization process is mathematically represented as:

$$\theta = \theta + \eta \nabla_\theta \mathbb{E}[R(P)], \tag{15}$$

where θ denotes the parameters of the prompt policy, η is the learning rate, and $R(P)$ represents the reward function tailored to task-specific Applications. Similarly, **SPARSETT** [596] employs sparse optimization techniques to retain only the most relevant tokens, ensuring computational efficiency. This refinement process is expressed as:

$$P_{\text{sparse}} = \underset{P}{\text{argmin}} \|P\|_0 \quad \text{subject to } R(P) \geq R_{\text{threshold}}. \tag{16}$$

The role of discrete prompts is visually depicted in **Figure 7 (Right)**, highlighting their integration into the input sequence to guide model outputs effectively. Building on these foundational structures, **adaptation strategies** refine how prompts are used to tailor pre-trained models for specific applications. **Task-specific adaptation** focuses on tailoring prompts for individual tasks to achieve high accuracy and efficiency. Techniques such as **Prompt Tuning** [434], **SK-Tuning** [529], **L-Tuning** [329], **XPrompt** [434], **IDPG** [724], **Arprompt** [683], **TPT** [717], and **SMoP** [94] fall into this category. For instance, **Prompt Tuning** introduces task-specific embeddings to the input sequence, enabling precise modulation of the model's outputs. This can be mathematically expressed as:

$$y = f(x, P_{\text{task-specific}}), \tag{17}$$

with x is the input, and $P_{\text{task-specific}}$ represents the learned prompt for the specific task. Iterative approaches like **IDPG (Iterative Dual Prompt Generation)** [724] refine prompts iteratively over multiple steps, expressed as:

$$P^{(t+1)} = P^{(t)} - \alpha \nabla_P \mathcal{L}(f(x, P^{(t)})), \tag{18}$$

where t represents the iteration step, and α is the learning rate.

5.4 Scaling PEFT

Scaling PEFT extends the utility of prompts to handle broader and more complex contexts. A notable advancement in this area is the **Propulsion concept** [327], which incorporates polynomial scaling to dynamically adjust the influence of prompt parameters. This mechanism allows for granular control over the model's sensitivity to input features and is mathematically defined as:

$$V_i' = [v_1 \odot z_1^k, v_2 \odot z_2^k, \ldots, v_s \odot z_s^k], \tag{19}$$

where v_i represents the input features, z_i are scaling parameters, k is the scaling exponent, and $\odot$ denotes element-wise multiplication. The Propulsion method's architecture is illustrated in **Figure 8** (right), showing its attention mechanism modification approach with selective parameter tuning.

5.5 Selective fine-tuning

Selective fine-tuning is an advanced model optimization technique designed to adapt pre-trained models to specific tasks by modifying only a carefully chosen subset of parameters while keeping the rest unchanged [325, 775, 250]. Unlike traditional fine-tuning, which updates all parameters in the model, selective fine-tuning focuses on parameters that are most relevant to the task. This targeted approach reduces computational costs, mitigates overfitting, and preserves the general-purpose knowledge embedded in the pre-trained model. By relying on principles such as parameter importance, sparsity, and structural organization, selective fine-tuning achieves a balance between efficiency and adaptability, making it an invaluable tool in modern machine learning.

The parameters of a pre-trained model, denoted as θ, are divided into two subsets: θ_s, the parameters selected for fine-tuning, and θ_f, the parameters that remain fixed. The selection of θ_s is guided by a criterion $C(\cdot)$, which evaluates the relevance of each parameter to the task. Parameters with relevance scores exceeding a threshold τ are included in θ_s, while the rest are assigned to θ_f. Mathematically, this can be expressed as:

$$\theta_s = \{\theta_i \mid C(\theta_i) \geq \tau\}, \quad \theta_f = \theta \setminus \theta_s.$$

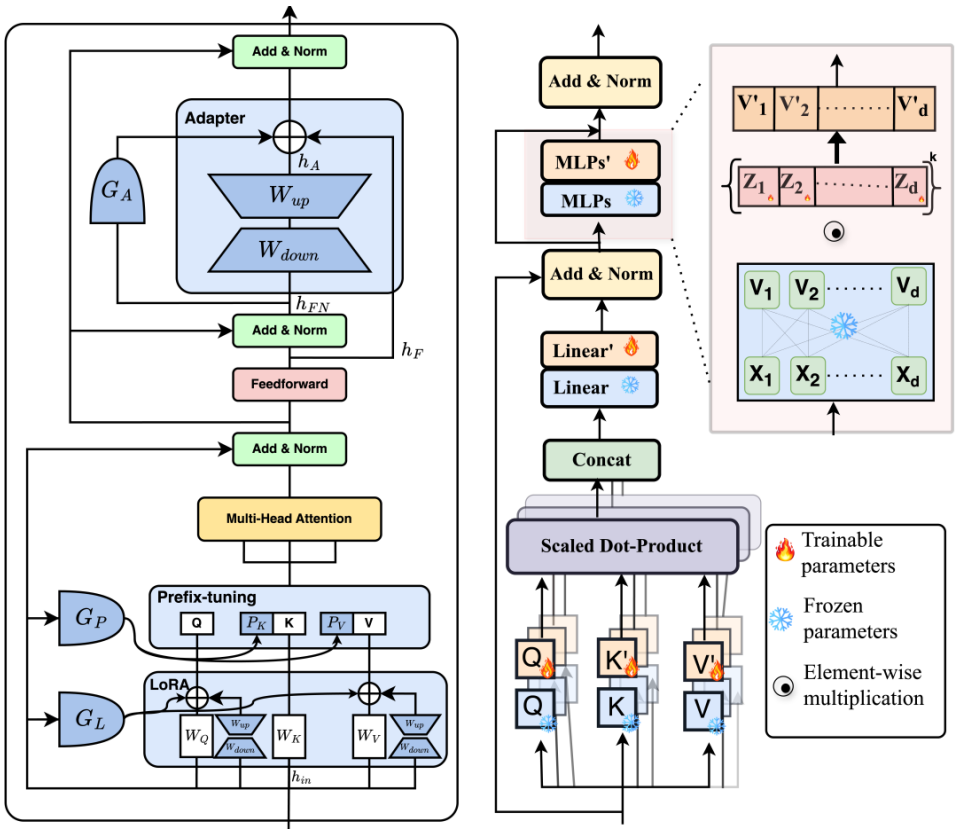


Figure 8: Architecture of a unified hybrid PEFT approach integrating multiple parameter-efficient fine-tuning strategies within a transformer block (left). Propulsion architecture showing trainable and frozen components. The design features scaled dot-product attention with selectively tunable parameters (orange) and frozen weights (blue), enabling efficient fine-tuning through element-wise multiplication of scaled vectors (right).

The optimization process then focuses on θ_s, while θ_f remains unchanged to preserve the pre-trained knowledge. This process is formalized as:

$$\underset{\theta_s}{\operatorname{argmin}} \mathcal{L}(f(x; \theta_s \cup \theta_f), y),$$

where $f(x;\theta)$ represents the model's output for input x, and $\mathcal{L}$ is the task-specific loss function. Selective fine-tuning often identifies important parameters based on their contribution to the task. Techniques such as Fisher information, gradient magnitudes, or sensitivity analysis are commonly used to measure parameter importance. Methods like **FishMask** [143], **Adafish** [250], **IST** [753], and **U-Diff Pruning** [199] exemplify this approach.

For example, **FishMask** [143] uses Fisher information to evaluate the importance of each parameter. The Fisher information for parameter θ_i is defined as:

$$\mathcal{I}(\theta_i) = \mathbb{E}\left[\left(\frac{\partial \log p(x;\theta)}{\partial \theta_i}\right)^2\right],$$

where $\mathcal{I}(\theta_i)$ quantifies the influence of θ_i on the model's predictions, and $\log p(x;\theta)$ is the log-likelihood. Parameters with high Fisher information are included in θ_s, as they have a significant impact on task performance. Similarly, **Adafish** [250] dynamically adjusts the selection criteria during training by analyzing gradient magnitudes, allowing the model to focus on parameters that are most relevant to the evolving task. As illustrated in **Figure 9**, this approach utilizes a dual-loop architecture to iteratively identify and select important parameters based on Fisher information, enabling more efficient task-specific adaptation.

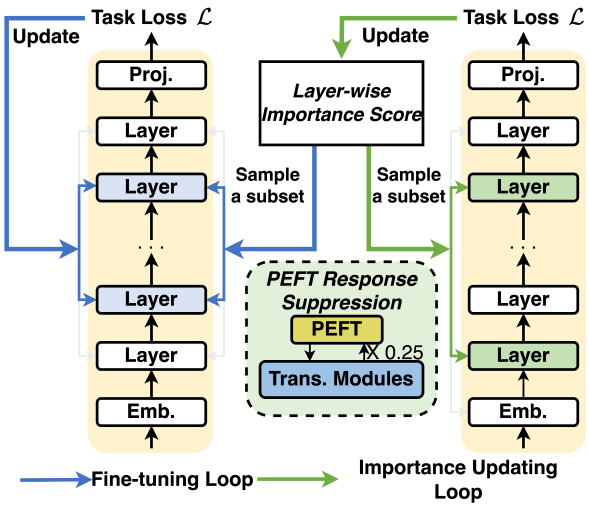


Figure 9: Parameter importance selection framework with dual optimization loops. Left: Fine-tuning loop updating model parameters based on task loss. Center: Layer-wise importance scoring mechanism with PEFT response suppression. Right: Importance updating loop that samples subset layers for targeted optimization based on their calculated importance scores.

Unstructured parameter selection focuses on selecting individual parameters independently of their groupings or positions within the model. This approach is commonly employed in sparsity-based techniques, such as **UBitFit** [343], **LT-SFT** [18], **Child-Tuning** [742], and **PaFi** [378].

In **Child-Tuning**, for instance, the relevance of each parameter is determined using gradient norms:

$$C(\theta_i) = \left\| \frac{\partial \mathcal{L}}{\partial \theta_i} \right\|,$$

where $C(\theta_i)$ is the importance score for parameter θ_i, and parameters with scores above a threshold τ are included in θ_s. This approach ensures that only the parameters that contribute significantly to the task are updated, enhancing both efficiency and performance. **Figure 10** illustrates the unstructured parameter selection process, showing how individual parameters are identified and updated to achieve task-specific optimization without disrupting the overall structure of the model.

Structured fine-tuning focuses on updating coherent groups of parameters, such as layers, attention heads, or blocks, rather than individual parameters. This approach is particularly effective for modular architectures like transformers, where parameters are hierarchically organized. Methods such as **RoCoFT** [325], **Far** [654], **Xattn Tuning** [186], **X-PEFT with Hard Masking** [338], and **SURM** [579] adopt this strategy.

In **X-PEFT with Hard Masking**, a binary mask M is applied to enforce structural constraints during fine-tuning:

$$\theta_s = M \odot \theta,$$

where $M \in \{0, 1\}^d$ represents the mask, $\odot$ denotes element-wise multiplication, and d is the number of parameters. This ensures that only critical components, such as specific layers or blocks, are updated, while the remaining parameters remain fixed. Similarly, **SURM** [579] applies domain-specific masking strategies to align fine-tuning with the structural requirements of the task. **Figure 11 (right)** illustrates different parameter update patterns in the **RoCoFT** [325] approach, showing how selective modification of rows or columns within weight matrices enables efficient fine-tuning while preserving model coherence.

5.6 Reparameterized PEFT

Reparameterized PEFT methods aim to optimize large-scale pre-trained models by introducing efficient, low-rank transformations that reduce trainable parameters while preserving task-specific performance [247, 403, 788, 181]. These methods can be broadly categorized into three groups:

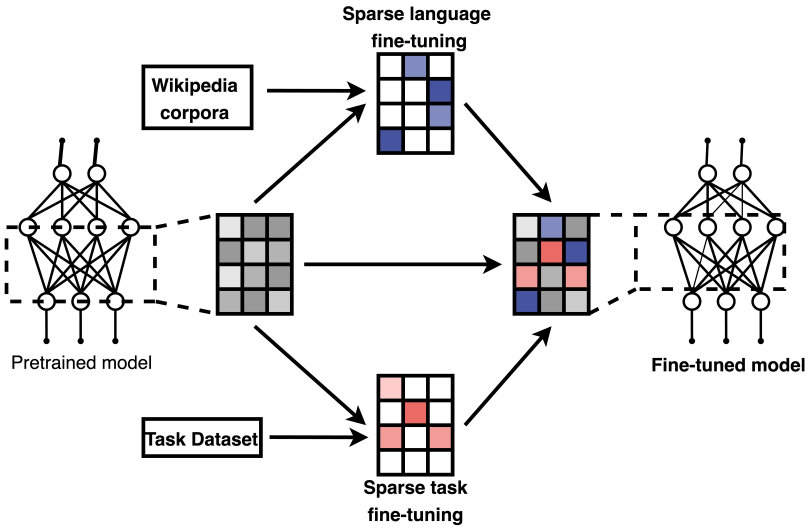


Figure 10: Sparse Fine-tuning Framework. This diagram illustrates the process of sparse parameter fine-tuning for language models. Starting with a pretrained model (left), the framework utilizes both Wikipedia corpora for sparse language fine-tuning (top path) and task-specific datasets for sparse task fine-tuning (bottom path). The selective parameter updates, represented by colored cells in the matrices, allow the fine-tuned model (right) to maintain general capabilities while adapting to specific tasks with minimal parameter changes. The blue cells represent language-related parameters, while red cells indicate task-specific parameters selected for updating.

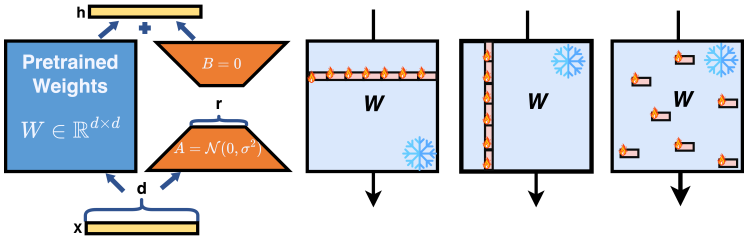


Figure 11: Visualization of LoRA and RoCoFT architectures. The first figure (left) illustrates the architecture of LoRA. The remaining figures depict the RoCoFT variants, including row-wise updates, column-wise updates, and random updates.

core low-rank decomposition techniques, adaptive and dynamic rank methods, and enhanced LoRA variants for specific tasks and fine-tuning efficiency. This paper provides a detailed overview of these methods and their applications in large-scale machine learning models. Reparameterized PEFT addresses the computational and memory constraints of fine-tuning large-scale models by introducing low-rank parameterization techniques. These approaches focus on reparameterizing the delta weight matrix (ΔW) into a low-dimensional form, significantly reducing the number of trainable parameters. The techniques can be classified into three main categories: **core low-rank decomposition**, **adaptive and dynamic rank methods**, and **enhanced LoRA variants** tailored for specific tasks.

5.6.1 Low-Rank Decomposition

The foundation of reparameterized PEFT lies in **low-rank decomposition**, where the parameter update matrix $\Delta W \in \mathbb{R}^{d \times d}$ is approximated as the product of two low-rank matrices, $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$, with $r \ll d$. Mathematically, this can be expressed as:

$$\Delta W \approx AB \tag{20}$$

This decomposition reduces the number of trainable parameters from d^2 to $2dr$, significantly lowering computational requirements. Methods such as **LoRA** [247], **Compactor** [295], **Intrinsic SAID** [7],

K Adaption [225], **DoRA** [403], and **LLEmbed** [392] build on this concept. For instance, LoRA constrains the updates to a low-rank subspace, while Compactor introduces sparsity-inducing priors for further efficiency. Intrinsic SAID optimizes updates using intrinsic dimensionality principles, and K Adaption dynamically tunes the rank r to align with task-specific requirements, enhancing flexibility. **Figure 11** illustrates the low-rank adaptation approach used in LoRA, demonstrating how the pretrained weight matrix $W \in \mathbb{R}^{d \times d}$ is combined with low-rank matrices A and B to form the final weight matrix $h = W \cdot x + BA \cdot x$, where B is initialized to zero and A is sampled from a Gaussian distribution.

This approach is particularly effective for large language models where the full fine-tuning of all parameters would be prohibitively expensive. By focusing training exclusively on the low-rank matrices A and B, LoRA achieves comparable performance to full fine-tuning while requiring only a fraction of the computational resources and storage requirements. The rank r serves as a hyperparameter that controls the trade-off between model capacity and training efficiency.

5.6.2 Dynamic Rank Methods

While core low-rank decomposition uses a fixed rank r, **adaptive and dynamic rank methods** adjust the rank during training to optimize performance and resource usage. Techniques such as **DyLoRA** [644] and **AdaLoRA** [788] dynamically modify r based on gradient information or layer sensitivity:

$$r_t = f(\|\nabla_t\|), \quad t \in \{1, \ldots, T\} \tag{21}$$

where $f(\cdot)$ is a function of the gradient norm $\|\nabla_t\|$, and T is the number of training steps. Similarly, **SLORA** [139] employs layer-wise rank scheduling, while **CapaBoost** [598] and **AutoLoRA** [790] automate rank selection using task-specific metrics. These methods introduce adaptability, ensuring efficient resource allocation and improved task performance. **Figure 12** illustrates the DyLoRA approach that dynamically adjusts low-rank updates through block-wise decomposition patterns, showcasing how parameter updates propagate through the model architecture while maintaining efficiency.

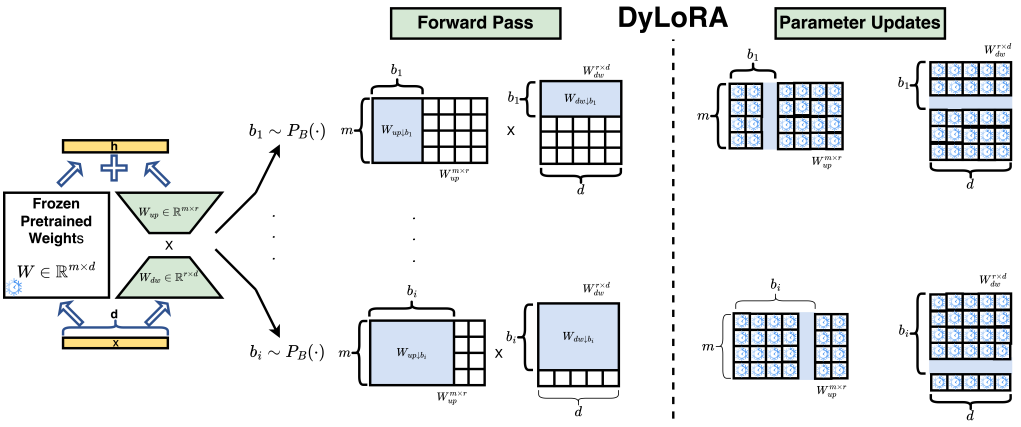


Figure 12: Visualization of DyLoRA (Dynamic Low-Rank Adaptation), which enhances standard LoRA by introducing dynamically sampled low-rank blocks. Left: Frozen pretrained weight matrix $W \in \mathbb{R}^{m \times d}$ with blocks W_{qp} and W_{kv}. Center: Forward pass showing how blocks $b_i \sim P_B(\cdot)$ are sampled and multiplied with corresponding weight matrices. Right: Parameter updates with dynamic allocation across matrix blocks, enabling more efficient fine-tuning by focusing updates where they provide the greatest benefit.

DyLoRA extends the standard LoRA framework by introducing block-wise dynamic parameter allocation, where update resources are distributed based on importance. Rather than applying uniform low-rank decomposition across all weight matrices, DyLoRA samples blocks b_i from a probability distribution $P_B(\cdot)$ and focuses on parameter updates in these regions. This targeted approach allows the model to concentrate computational resources where they will be most impactful, further reducing training overhead while maintaining or even improving adaptation quality compared to static low-rank methods.

5.6.3 LoRA Variants

Building on the foundational low-rank framework, enhanced variants of LoRA address domain-specific challenges and improve fine-tuning efficiency. Methods like **Laplace LoRA** [746], **LoRA Dropout** [386], and **Predict LoRA** [460] introduce regularization and dropout techniques to mitigate overfitting. For example, Laplace LoRA augments the decomposition with a regularization term:

$$\Delta W \approx AB + \lambda I, \quad \lambda > 0 \tag{22}$$

where λ controls the regularization strength. **LoRA++** [215], **MoSLoRA** [720], and **LoRA for Continual Learning** [93] are tailored to sequential learning tasks, effectively preventing catastrophic forgetting. On the other hand, **Trans-LoRA** [685] and **RoseLoRA** [675] extend LoRA to transfer learning scenarios, adapting pre-trained models to new domains through task-specific subspaces.

Further innovations address fine-tuning challenges in low-resource settings. **LoRA for Few-Shot Learning** [777], **SVDQUANT** [370], and **Variational LoRA IVON** [114] enhance efficiency through quantization and probabilistic modeling. For example, SVDQUANT performs singular value decomposition (SVD) followed by quantization, while Variational LoRA incorporates Bayesian principles to account for uncertainty:

$$p(W|\mathcal{D}) \propto p(\mathcal{D}|W)p(W), \quad W = W_0 + AB \tag{23}$$

where W_0 is the original weight matrix, and AB is the low-rank update. Ensemble methods, such as **MoeLoRA** [401], **MoLoRA** [773], and **MixLoRA** [359], integrate multiple low-rank models to improve robustness and generalization. Finally, **LoRA Hub** [255] consolidates diverse PEFT strategies into a unified framework, facilitating their application across varied tasks. **Figure 13** illustrates dropout regularization techniques applied to LoRA and AdaLoRA, showcasing how selective parameter dropping during training enhances model robustness and prevents overfitting.

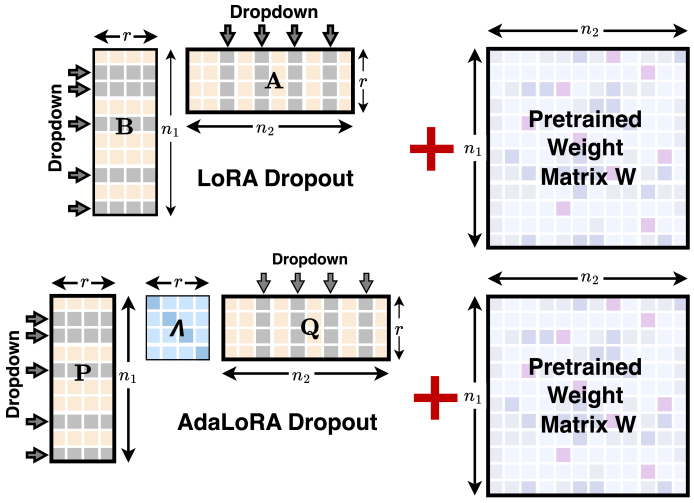


Figure 13: Comparison of dropout regularization strategies for low-rank adaptation methods: Top - LoRA Dropout applies structured dropout to low-rank matrices $A \in \mathbb{R}^{r \times n_2}$ and $B \in \mathbb{R}^{n_1 \times r}$ during training, randomly dropping elements along both row and column dimensions; Bottom - AdaLoRA Dropout extends this concept with matrices P, Λ, and Q, providing more flexible regularization patterns while maintaining the efficiency benefits of low-rank decomposition. Both approaches preserve the pretrained weight matrix $W \in \mathbb{R}^{n_1 \times n_2}$ while selectively regularizing the update components.

These dropout techniques strategically disable portions of the low-rank matrices during training, which serves multiple purposes: preventing co-adaptation of weight updates, improving generalization by creating an implicit ensemble effect, and further reducing computational demands. In LoRA Dropout, elements in matrices A and B are randomly zeroed according to dropout patterns along both dimensions, while AdaLoRA Dropout implements a more sophisticated approach with its three-matrix decomposition. These regularization methods are particularly valuable for scenarios where the fine-tuning dataset is limited, as they help prevent the model from simply memorizing

training examples while maintaining the parameter efficiency that makes low-rank adaptation methods attractive.

5.7 Hybrid PEFT

Hybrid approaches in PEFT combine multiple fine-tuning strategies, such as **LoRA**, **adapters**, and **prompt-tuning**, into a unified framework to leverage the strengths of each method. By integrating these techniques, hybrid approaches provide flexibility, adaptability, and robustness across diverse tasks. These methods dynamically determine the most suitable combination of strategies to optimize performance while maintaining efficiency.

For example, the **MAM Adapter** [217] incorporates memory components into adapters, allowing task-specific information to be stored and retrieved, thereby enhancing the model's ability to specialize in different tasks. Similarly, **UniPELT** [447] (Unified Parameter-Efficient Language Tuning) integrates **LoRA**, **prefix-tuning**, and **adapters** within a single framework, enabling the model to switch dynamically between strategies depending on the task. Another prominent method, **RoSA** (Rank-Ordered Subspace Adaptation) [490], prioritizes the most significant subspaces of parameters for fine-tuning. This is achieved by rank-ordering parameters and selecting the top-ranked ones for updates:

$$\theta_s = \{\theta_i \,|\, \text{rank}(\theta_i) \leq k\},$$

where k is the threshold for the top-ranked parameters.

Hybrid approaches often use a weighted combination of parameter updates:

$$\theta_{\text{hybrid}} = \sum_{i=1}^{n} \alpha_i \theta_i,$$

where α_i represents the dynamically adjusted weight for the i-th strategy, and n is the number of integrated strategies. This framework allows hybrid methods to balance computational efficiency with task-specific adaptability.

Additional methods such as **S4** [75], **NOAH** [795], **Auto PEFT** [813], **LLM Adapter** [254], **SH-PEFT** [400], **Hyper PELT** [797], and **Hydra** [313] extend the versatility of hybrid approaches by automating strategy selection, focusing on structured sparsity, or incorporating multi-headed designs for enhanced flexibility. **Figure 8 (left)** illustrates the architecture of a unified hybrid approach, demonstrating how multiple parameter-efficient fine-tuning methods can be integrated within a single transformer block.

This unified architecture elegantly combines the strengths of multiple PEFT approaches: Adapters provide sequential transformation through bottleneck architectures, Prefix-tuning prepends learnable vectors to modify attention patterns, and LoRA applies low-rank updates to weight matrices. The inclusion of gating mechanisms (G_A, G_P, G_L) enables the model to dynamically weight the contribution of each method based on task requirements. This hybrid design achieves superior performance by leveraging complementary benefits: Adapters excel at capturing task-specific transformations, Prefix-tuning provides efficient context modification, and LoRA delivers parameter-efficient weight adjustments. The unified approach not only improves task performance but also enhances transfer learning capabilities across diverse domains while maintaining the parameter efficiency that makes PEFT methods attractive for resource-constrained environments.

5.7.1 MoE-Based

An emerging class of **Mixture-of-Experts (MoE)-based** PEFT methods extends low-rank adaptation by incorporating expert routing mechanisms that dynamically select or combine multiple low-rank modules during training or inference. These methods aim to improve model specialization and generalization across diverse tasks while maintaining parameter efficiency. Formally, the update matrix ΔW is expressed as a weighted sum of expert-specific low-rank transformations:

$$\Delta W = \sum_{i=1}^{n} \alpha_i A_i B_i, \quad \sum_{i=1}^{n} \alpha_i = 1 \tag{24}$$

where $A_i B_i$ represents the i-th low-rank expert, and α_i is a gating coefficient determined by the MoE router. This formulation enables input-dependent specialization by activating only the most

relevant subset of experts, reducing computational overhead while enhancing adaptability. Several MoE-based methods have been proposed to leverage this framework. **MoE LoRA** [401] introduces a learned gating mechanism to select among multiple LoRA experts, facilitating dynamic specialization across inputs. **MixLoRA** [359] combines several LoRA modules through task-aware mixture weights, improving robustness and domain generalization. **MoLoRA** [773] routes tokens to different LoRA experts at each transformer layer, enabling fine-grained control over parameter updates. **MOA** (Mixture of Adaptations) [169] generalizes this idea by integrating multiple adaptation strategies—such as LoRA, adapters, and prefix tuning—within a unified routing framework. Finally, **MoLE** (Mixture of Low-rank Experts) [720] consolidates several low-rank experts and selects them dynamically based on input features, enhancing scalability and performance in multi-task and low-resource settings. **Figure 14** illustrates the taxonomy and relationships among MoE-based PEFT methods, highlighting how they extend traditional low-rank approaches with modular, expert-driven architectures to support efficient, task-adaptive fine-tuning.

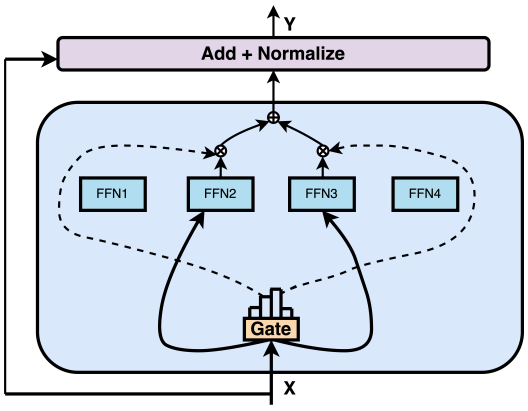


Figure 14: Illustration of a Mixture-of-Experts (MoE) feedforward network layer with gated expert selection. The input X is routed to multiple feedforward sub-networks (experts), labeled as FFN1 through FFN4. A gating mechanism computes routing weights to determine which subset of experts to activate for a given input. In this example, two experts (FFN2 and FFN3) are selected and their outputs are weighted and combined. The result is added to the residual connection and passed through a normalization layer to produce the final output Y. This structure enables conditional computation, enhancing model capacity while maintaining computational efficiency.

6 Experiments

6.1 GLUE Benchmark Performance Comparison:

We conducted a comprehensive evaluation of various PEFT methods across the General Language Understanding Evaluation (GLUE) benchmark [660] using both RoBERTa$_{\text{Base}}$ and RoBERTa$_{\text{Large}}$ models [413]. The GLUE tasks include a diverse range of linguistic challenges, such as single-sentence classification (CoLA and SST-2), sentence-pair classification (MRPC, QQP, MNLI, QNLI, and RTE), and regression-based semantic similarity (STS-B). For evaluation, we employed standard metrics: Matthews Correlation Coefficient (MCC) for CoLA, accuracy for SST-2, accuracy and F1 score for MRPC and QQP, Pearson and Spearman correlations for STS-B, and accuracy for MNLI, QNLI, and RTE. This analysis aimed to determine the trade-off between model performance and parameter efficiency across established and novel PEFT techniques, including the recently introduced SK-Tuning method.

Table 1 presents a comprehensive evaluation of various parameter-efficient fine-tuning (PEFT) methods on the GLUE benchmark using RoBERTa$_{\text{Base}}$ and RoBERTa$_{\text{Large}}$ models. Full fine-tuning (FT), which updates all model parameters, consistently yields strong performance across all tasks but incurs high computational and memory costs. It serves as a performance upper bound for assessing the efficiency of PEFT techniques.

PEFT Method	# TTPs	CoLA	SST2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
RoBERTa$_{\text{Base}}$									
FT	124.6M	59.84	92.89	85.24/88.18	90.48/90.16	90.18/87.02	86.27	91.17	72.43
Adapter^{S}	7.41M	61.53	94.11	89.81/90.85	90.25/90.09	89.81/86.90	86.27	92.06	73.56
Prompt tuning	0.61M	49.37	92.09	70.83/81.72	82.44/83.11	82.99/78.35	80.57	80.03	58.12
Prefix-tuning	0.96M	59.31	93.81	84.25/85.03	88.48/88.32	87.75/84.09	85.21	90.77	54.51
(IA)^3	0.66M	58.58	93.92	83.00/85.52	90.30/90.32	87.99/84.10	83.95	90.88	71.12
BitFit	0.083M	61.32	93.12	87.22/88.41	90.34/90.27	88.12/84.11	84.64	91.09	77.98
LoRA	0.89M	60.09	93.31	86.50/88.68	90.66/90.47	88.83/85.21	86.54	92.02	74.92
AdaLoRA	1.03M	59.82	93.92	86.49/88.03	90.83/90.73	88.58/84.98	86.26	91.43	76.04
MAM Adapter	1.78M	58.34	94.24	87.31/88.21	90.74/90.42	88.31/83.20	86.63	90.19	72.62
PROPETL$_{\text{Adapter}}$	1.87M	64.24	93.85	87.15/87.82	90.33/90.64	89.22/85.79	86.49	91.56	75.54
PROPETL$_{\text{Prefix}}$	10.49M	60.11	93.63	86.73/87.98	90.30/90.19	88.54/85.05	86.22	91.51	63.31
PROPETL$_{\text{LoRA}}$	1.77M	57.94	94.11	87.42/88.87	90.66/90.35	88.90/85.55	86.83	92.04	67.39
MoSLoRA	1.67M	60.57	93.95	86.74/87.98	90.05/89.43	88.76/85.62	87.84	90.60	75.10
LoRA-XS	0.26M	58.49	93.19	86.65/87.49	89.60/89.33	87.13/84.31	85.34	90.42	76.24
VeRA	0.043M	60.35	93.89	86.01/87.88	89.27/89.41	87.88/85.65	85.64	90.22	75.32
LoRAFA	0.44M	60.49	93.65	88.18/89.98	90.70/90.66	88.90/85.46	86.11	91.42	76.11
SFT	0.90M	64.45	94.28	87.74/88.64	89.37/89.12	87.24/85.11	86.64	92.11	78.42
Diff Pruning	1.24M	62.45	93.77	88.00/89.21	89.72/90.02	88.62/85.54	85.32	92.14	77.90
RoCoFT$_{\text{1-Row}}$	0.083M	60.18	94.06	87.74/88.48	90.70/90.47	88.49/85.35	85.23	90.70	76.61
RoCoFT$_{\text{3-Row}}$	0.249M	63.53	94.92	87.71/88.74	90.89/90.49	88.97/85.80	86.73	92.12	78.31
RoCoFT$_{\text{1-Column}}$	0.083M	60.32	93.88	88.38/89.78	90.23/90.14	88.46/85.84	85.35	90.58	76.74
RoCoFT$_{\text{3-Column}}$	0.249M	62.95	94.69	89.18/90.94	90.85/90.45	88.86/85.38	86.76	91.89	75.21
Propulsion$_{\text{1-Row}}$	0.086M	61.76	93.18	89.34/85.99	90.37/89.92	89.11/86.53	86.41	91.79	75.66
Propulsion$_{\text{3-Row}}$	0.258M	63.21	94.35	87.28/86.12	90.29/90.04	89.42/86.84	87.12	91.56	76.92
Propulsion$_{\text{Attn}}$	0.028M	58.51	92.03	89.01/85.14	89.36/89.96	86.73/84.80	85.13	89.89	75.02
SK-Tuning (Prompt)	0.60M	60.21	93.88	89.73/92.47	91.30/90.19	87.83/85.82	86.24	92.60	76.91
SK-Tuning (Prefix)	0.84M	61.83	93.72	87.21/88.04	90.11/89.92	88.67/87.12	85.83	92.09	75.32
RoBERTa$_{\text{Large}}$									
FT	355.3M	65.78	95.50	92.22/94.28	91.74/91.96	90.83/88.68	89.21	93.19	81.40
Adapter^{S}	19.77M	65.33	96.37	89.88/90.23	92.58/92.42	91.19/87.11	91.00	94.31	85.25
Prompt-tuning	1.07M	61.13	94.61	73.04/76.29	78.51/78.99	80.74/75.16	68.15	89.13	60.29
Prefix-tuning	2.03M	59.01	95.76	88.24/89.37	90.92/91.07	88.88/85.45	89.30	93.32	74.01
(IA)^3	1.22M	61.15	94.61	86.45/87.53	92.22/86.25	89.45/86.25	88.63	94.25	81.23
BitFit	0.222M	67.01	96.10	90.93/92.13	91.93/93.38	89.48/86.43	89.98	94.47	87.73
LoRA	1.84M	64.47	96.67	87.50/88.19	91.66/91.44	90.15/86.91	90.76	95.00	79.78
AdaLoRA	2.23M	65.85	94.95	89.46/90.34	92.05/91.80	89.60/86.30	90.36	94.62	77.98
MAM Adapter	4.20M	67.39	95.81	90.12/92.07	92.44/92.18	90.87/86.65	90.62	94.31	86.62
PROPETL$_{\text{Adapter}}$	5.40M	65.55	96.27	89.71/91.15	91.92/91.67	90.67/87.74	91.37	94.80	87.69
PROPETL$_{\text{Prefix}}$	26.85M	62.24	96.17	90.04/91.92	90.70/90.49	89.30/86.30	90.33	94.73	79.71
PROPETL$_{\text{LoRA}}$	4.19M	61.90	95.93	87.31/89.87	91.66/91.38	90.93/88.05	90.53	94.93	83.57
MoSLoRA	3.23M	67.27	96.17	89.96/92.67	90.97/91.72	90.12/87.68	90.29	94.73	82.41
RoCoFT$_{\text{1-Row}}$	0.222M	65.70	96.63	89.97/90.79	91.81/92.07	90.17/86.15	90.73	94.20	85.31
RoCoFT$_{\text{3-Row}}$	0.666M	67.39	96.69	91.05/92.19	92.10/92.10	90.82/86.11	90.98	94.85	87.83
RoCoFT$_{\text{1-Column}}$	0.222M	64.89	96.60	89.12/90.24	91.96/92.10	90.17/85.83	90.81	94.17	85.71
RoCoFT$_{\text{3-Column}}$	0.666M	67.18	96.67	89.88/91.47	92.52/92.31	91.38/87.12	91.13	94.85	87.82
Propulsion$_{\text{1-Row}}$	0.225M	64.53	95.10	90.47/88.85	91.78/91.58	92.26/88.91	90.52	95.34	85.30
Propulsion$_{\text{3-Row}}$	0.675M	67.12	96.68	91.15/92.07	91.68/91.81	91.96/87.84	91.42	95.12	88.28
Propulsion$_{\text{Attn}}$	0.073M	62.31	94.02	89.78/87.95	90.16/90.86	88.02/86.19	89.54	94.00	83.07
SK-Tuning (Prompt)	1.02M	67.13	96.43	91.10/93.22	90.54/90.11	92.10/88.73	90.42	95.42	87.11
SK-Tuning (Prefix)	1.94M	66.33	96.08	90.96/93.09	91.87/90.68	90.23/87.93	89.97	96.10	86.99

Table 1: RoBERTa models performance on GLUE tasks: Metrics used are MCC for CoLA, accuracy for SST-2, accuracy/F1 score for MRPC and QQP, Pearson/Spearman correlations for STS-B, and accuracy for MNLI, QNLI, and RTE.

Among the PEFT baselines, **Adapter^{S}**, **BitFit**, and **LoRA** perform remarkably well. For example, BitFit (0.083M parameters) achieves 77.98% on RTE and 93.12% on SST-2, rivaling full fine-tuning. LoRA (0.89M) consistently outperforms most early PEFT methods and even FT in certain tasks, such as MNLI and QNLI. Adapter^{S} also demonstrates strong performance, particularly with RoBERTa$_{\text{Large}}$, scoring 96.37% on SST-2 and 85.25% on RTE.

Prompt-tuning and **Prefix-tuning**, while highly parameter-efficient (under 1M parameters), generally underperform on tasks requiring fine-grained semantic understanding, such as MRPC, STS-B, and RTE. This highlights their limited expressive capacity despite their minimal footprint.

Advanced LoRA-based methods such as **AdaLoRA**, **MoSLoRA**, and **LoRAFA** improve performance further. AdaLoRA, for instance, achieves 76.04% on RTE and 90.83% on STS-B with RoBERTa$_{\text{Base}}$, indicating the benefit of adaptive low-rank decompositions. MoSLoRA (1.67M) performs particularly well on MNLI (87.84%) and QNLI (90.60%), suggesting it captures diverse token-level information more effectively.

The **RoCoFT** and **Propulsion** families deliver better results among compact methods. RoCoFT$_{\text{3-Row}}$ and RoCoFT$_{\text{3-Column}}$ attain scores close to or exceeding FT on several tasks. Notably, RoCoFT$_{\text{3-Row}}$

reaches 78.31% on RTE and 94.92% on SST-2, with only 0.249M parameters. Similarly, Propulsion$_{\text{3-Row}}$ matches or surpasses strong baselines, achieving 76.92% on RTE and 94.35% on SST-2 with just 0.258M parameters. Even ultra-light versions like Propulsion$_{\text{Attn}}$ (0.028M) score competitively on tasks like STS-B and MRPC.

SK-Tuning, a recent method that integrates semantic knowledge into prompt and prefix tuning, demonstrates robust performance. SK-Tuning (Prompt) with 0.60M parameters achieves 92.60% on QNLI and 76.91% on RTE, outperforming traditional prompt-based approaches. Its prefix variant also performs well across all tasks, suggesting that semantically-aware prompting offers a powerful alternative for low-resource fine-tuning.

Finally, comparing across model sizes, PEFT methods applied to RoBERTa$_{\text{Large}}$ typically outperform their RoBERTa$_{\text{Base}}$ counterparts by a significant margin. For instance, RoCoFT$_{\text{3-Row}}$ achieves 87.83% on RTE with RoBERTa$_{\text{Large}}$, compared to 78.31% with RoBERTa$_{\text{Base}}$, highlighting the scaling benefits of PEFT with larger backbones.

In summary, modern PEFT methods—particularly LoRA-based variants, RoCoFT, Propulsion, and SK-Tuning—approach or even surpass full fine-tuning performance on many GLUE tasks while drastically reducing the number of updated parameters. This makes them highly attractive for efficient and scalable deployment of large language models in both academic and production settings.

6.2 LLM Reasoning PEFT Comparison :

Method	# TTPs	BoolQ	PIQA	SIQA	H.Sw.	W.Gra.	ARCe	ARCc	OBQA	M.Ar.	G.8K	A.S.	Sing.Eq	S.MP
BLOOM$_{\text{7B}}$														
Prefix	33.37M	58.53	62.24	65.41	48.32	66.63	68.13	49.32	63.51	78.41	66.45	67.52	66.94	49.10
AdaLoRA	24.88M	66.94	74.68	72.49	55.89	68.30	73.21	56.59	72.85	79.43	70.25	68.93	70.93	53.89
$(IA)^3$	19.34M	63.30	73.33	71.01	52.50	71.60	69.45	54.14	68.60	78.90	71.17	70.33	70.84	53.95
LoRA	24.22M	65.89	73.92	73.33	56.65	71.39	73.46	57.15	72.31	79.50	70.93	70.90	70.59	54.85
RoCoFT$_{\text{3-Row}}$	13.37M	66.33	74.53	73.56	56.60	72.14	73.29	57.48	72.92	79.76	70.94	70.95	70.90	54.42
RoCoFT$_{\text{3-Column}}$	13.37M	66.34	74.64	71.12	55.93	72.50	73.11	57.19	72.90	79.72	71.05	70.88	70.76	54.38
Propulsion	13.37M	66.38	74.63	73.62	57.25	72.33	73.09	57.61	73.12	79.36	70.95	70.92	71.22	53.52
GPT-J$_{\text{6B}}$														
Prefix	27.83M	62.28	65.04	67.72	44.15	63.71	63.59	46.47	58.31	83.12	67.44	75.25	78.46	49.12
AdaLoRA	20.77M	65.19	67.58	71.22	45.16	66.03	64.10	47.75	63.92	88.51	73.45	80.21	83.03	56.14
$(IA)^3$	16.61M	63.17	68.51	68.97	45.79	66.06	62.42	45.32	65.42	89.51	72.04	80.50	81.50	55.43
LoRA	20.02M	65.50	68.63	69.46	45.60	66.80	65.56	46.81	63.82	88.30	72.82	80.60	81.24	56.73
RoCoFT$_{\text{3-Row}}$	11.62M	65.92	68.53	69.90	45.97	66.87	64.91	45.12	65.07	89.45	72.80	80.45	82.12	56.79
RoCoFT$_{\text{3-Column}}$	11.62M	65.12	68.22	69.96	45.98	66.78	64.89	45.70	64.81	89.74	72.24	80.23	82.61	56.70
Propulsion	11.62M	65.97	68.05	69.96	45.99	66.18	64.45	46.95	64.56	89.19	72.82	81.41	81.42	56.68
LLama-2$_{\text{7B}}$														
Prefix	33.53M	67.33	79.46	75.80	76.04	72.11	71.67	57.33	69.98	84.18	68.47	81.04	80.00	52.17
AdaLoRA	24.90M	67.03	80.69	76.06	88.85	76.47	76.50	61.36	74.22	89.81	77.07	86.70	83.01	60.25
$(IA)^3$	19.42M	69.02	78.10	78.00	87.57	76.78	75.48	60.54	74.02	90.20	76.13	86.55	83.70	59.16
LoRA	24.30M	69.89	79.37	76.15	88.86	77.54	76.54	60.55	74.63	90.13	75.68	84.67	82.14	59.94
RoCoFT$_{\text{3-Row}}$	13.47M	69.36	80.01	78.09	87.28	76.73	76.46	60.55	75.55	90.37	76.12	86.66	82.75	59.92
RoCoFT$_{\text{3-Column}}$	13.47M	69.32	80.08	77.99	87.46	76.41	76.46	60.59	74.90	90.42	77.35	86.16	82.48	60.35
Propulsion	13.47M	68.99	79.47	77.02	76.73	76.06	76.64	61.29	74.76	90.21	77.57	85.63	82.60	60.51
LLama-2$_{\text{13B}}$														
Prefix	61.97M	68.38	80.99	77.80	80.00	76.35	77.62	61.32	72.94	87.22	71.09	84.09	81.28	58.25
AdaLoRA	45.04M	71.71	82.55	78.88	91.60	83.01	83.04	67.33	81.76	90.55	80.19	87.00	87.10	66.03
$(IA)^3$	36.02M	71.39	83.33	78.32	92.40	83.24	83.34	66.43	80.99	91.88	79.24	88.16	87.08	67.03
LoRA	44.94M	71.19	83.99	79.15	91.86	83.24	83.35	67.05	81.37	91.27	78.90	86.89	86.07	65.85
RoCoFT$_{\text{3-Row}}$	24.88M	71.46	83.32	79.54	91.86	83.22	83.65	67.12	81.54	90.69	79.70	88.24	87.28	66.60
RoCoFT$_{\text{3-Column}}$	24.88M	71.44	83.52	79.50	91.84	83.20	83.39	67.06	81.73	91.46	79.63	88.11	87.58	66.63
Propulsion	24.88M	71.93	83.12	79.01	90.73	83.60	83.44	67.64	81.38	90.91	78.71	87.64	87.11	66.67

Table 2: Accuracy comparison of commonsense and mathematical reasoning performance across different PEFT methods using LLMs.

Table 2 provides a detailed comparison of PEFT methods on a diverse set of reasoning tasks—including commonsense (BoolQ [103], PIQA [47], SIQA [568], HellaSwag [778], WinoGrande [563], ARCe [105], ARCc [105], OBQA [466]) and mathematical/logical reasoning (MathQA [15], GSM8K [111], Arithmetic Sequence (A.S.) [240], SVAMP [513], and SingleEq [321]). The analysis spans four large language models (LLMs): BLOOMZ$_{\text{7B}}$ [712], GPT-J$_{\text{6B}}$ [663], LLaMA-2$_{\text{7B}}$ [633], and LLaMA-2$_{\text{13B}}$ [633]. We observe consistent patterns in performance improvements across PEFT methods and models.

Across all LLMs, full prefix tuning serves as a baseline and generally underperforms compared to more advanced PEFT methods, despite using a relatively large number of trainable parameters (e.g.,

61.97M for LLaMA-2_{13B}). In contrast, **AdaLoRA**, **(IA)^3**, and **LoRA** deliver substantial gains in reasoning benchmarks while reducing the parameter budget by 25–40%. Notably, AdaLoRA achieves robust results across most tasks, particularly with LLaMA-2_{13B}, scoring 91.60% on HellaSwag, 83.01% on WinoGrande, and 66.03% on SingleEq.

RoCoFT and **Propulsion**, both low-rank, structure-aware fine-tuning strategies, consistently match or outperform other PEFT baselines with significantly fewer trainable parameters. For example, RoCoFT$_{3\text{-Row}}$ and Propulsion, each with only 13.37M parameters on BLOOMZ$_{7B}$, outperform both AdaLoRA and LoRA on multiple tasks such as PIQA (74.63%) and SIQA (73.62%), while maintaining comparable scores on MathQA, GSM8K, and SVAMP. On GPT-J$_{6B}$, RoCoFT and Propulsion similarly demonstrate improvements over LoRA, especially on arithmetic and symbolic reasoning benchmarks like AquaRat and SVAMP, reflecting their potential to capture deeper reasoning patterns with minimal parameter cost.

With LLaMA-2_{7B}, performance increases across the board. LoRA and RoCoFT$_{3\text{-Row}}$ show strong results on difficult commonsense tasks such as HellaSwag (88.86% and 87.28%) and ARCc (60.55% for both). Meanwhile, Propulsion achieves near-competitive results (e.g., 77.57% on GSM8K) while maintaining efficiency. This further supports that structural PEFT methods can scale well to larger models without compromising generalization ability.

The LLaMA-2_{13B} model yields the highest overall accuracy, with all PEFT methods outperforming their smaller model counterparts. RoCoFT$_{3\text{-Row}}$ and Propulsion reach peak performance on SIQA (79.54%), HellaSwag (91.86%), and OBQA (81.54%), matching or exceeding AdaLoRA despite requiring nearly half the trainable parameters. For mathematical reasoning tasks like GSM8K and AquaRat, (IA)^3 and Propulsion offer strong performance, indicating that selective structural adaptation helps retain precision in arithmetic operations and symbolic pattern generalization.

In summary, while classical methods such as LoRA and AdaLoRA continue to perform strongly, newer PEFT techniques like RoCoFT and Propulsion demonstrate impressive performance-per-parameter efficiency across a wide range of reasoning tasks and model sizes. These approaches not only reduce computational costs but also scale robustly with model size, making them ideal for fine-tuning large LLMs on complex reasoning domains in real-world applications.

7 Applications

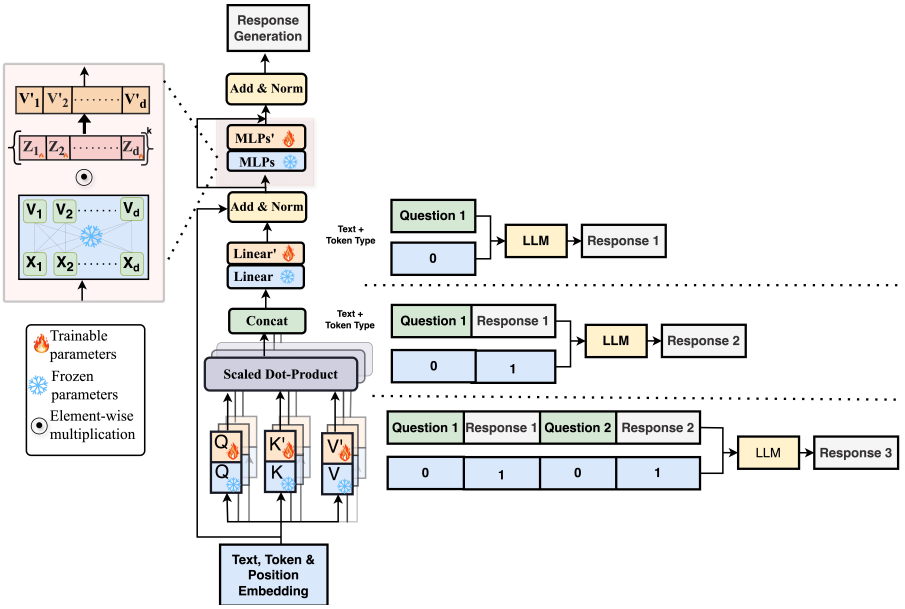


Figure 15: The figure depicts a Propulsion into ChatLLM framework. Token-type embeddings handle multi-turn dialogue, while a reparameterization mechanism modulates activations via element-wise multiplication for efficient adaptation in tasks like instruction tuning and personalization.

7.1 PEFT in NLP

PEFT techniques have been widely adopted across a range of NLP applications, offering an efficient way to adapt large language models (LLMs) to task-specific Applications without incurring the high cost of full fine-tuning. In text classification tasks such as sentiment analysis [699, 456, 615, 459, 60, 262, 390, 135], spam detection [117, 283, 449, 662, 731, 666], and topic categorization [600, 292, 815, 812, 533], PEFT methods allow models to be fine-tuned on relatively small labeled datasets while retaining strong performance, particularly in low-resource or domain-specific settings. For sequence generation tasks like text summarization [411, 704, 155, 622, 179, 246, 2], information extraction [528], and machine translation [669, 419, 320, 263, 303, 722], PEFT enables the model to adapt to domain-specific vocabulary and style, achieving competitive results with a fraction of the training parameters.

In dialogue systems, especially in multi-turn chat applications [367, 792, 339, 293, 677], PEFT plays a crucial role by enabling LLMs to handle evolving context and intent across conversation turns. Notably, PEFT has been integrated into frameworks like ChatLLM [326, 213, 553], where it supports efficient training and deployment of chat models by modifying only selected parameters—such as adapters or token embeddings—while keeping the core model frozen. This allows for rapid customization to different user personas, use-cases, or industries (e.g., healthcare, customer support) without retraining the entire model.

PEFT is also instrumental in instruction tuning and prompt-based learning, where models are aligned to follow specific instructions or exhibit desired behavioral traits. In few-shot and zero-shot scenarios, PEFT enables effective model adaptation with very limited examples. Furthermore, in multi-task and cross-lingual NLP setups [85, 553, 213, 116], PEFT allows a single LLM to be fine-tuned for multiple tasks or languages using task-specific adapters, thus promoting parameter sharing and memory efficiency.

An overview of this architecture is shown in **Figure 15**, where the Propulsion method [327] is integrated into the ChatLLM framework [326, 213] for training chat models with LLMs.

Additionally, we present a comprehensive summary of PEFT methods applied in NLP tasks across various LLMs and datasets in Tables 5, 6, 7, 8, 9, 10, 11, and 12.

Overall, PEFT has become a cornerstone of practical NLP development, making the deployment of powerful LLMs feasible for a wide variety of real-world applications and resource-constrained environments.

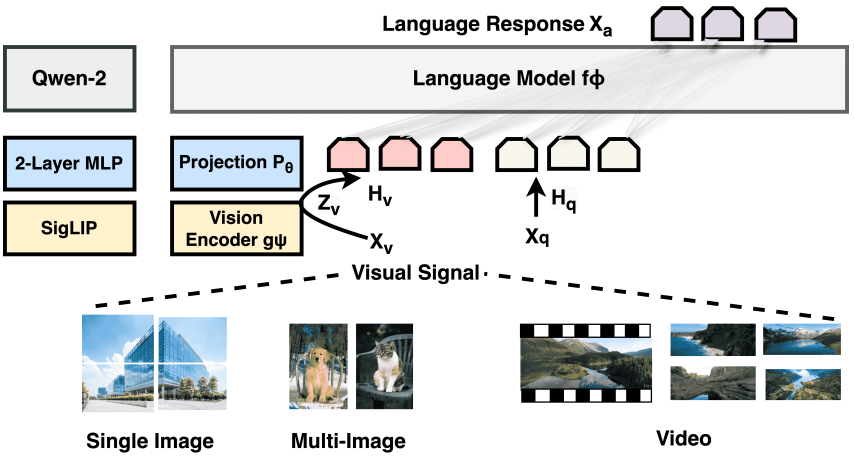


Figure 16: Diagram Description: The figure demonstrates a pipeline where vision embeddings (from SigLIP) are projected into the LLM input space via trainable layers. The model processes varying visual contexts (e.g., a cat or a natural scene) to produce structured language outputs, suitable for captioning, summarization, and vision-grounded reasoning.

7.2 PEFT in Vision

PEFT has also gained significant traction in computer vision, where large vision models—such as Vision Transformers (ViTs) [146] and large-scale convolutional neural networks (CNNs) [441, 326, 70] —are increasingly being adapted to diverse downstream tasks. In image classification [686, 420, 77, 543], PEFT enables efficient domain adaptation by allowing models pretrained on large datasets like ImageNet [672] to be fine-tuned on task-specific or domain-specific datasets (e.g., medical imaging [296, 245, 588, 40, 180, 612], satellite imagery [50, 33, 123, 268, 544]) with minimal additional parameters. This is particularly valuable in cases where data is scarce or computational resources are limited.

In object detection [824, 804, 16, 509] and instance segmentation [301], PEFT techniques such as adapter-based tuning or bias tuning have been applied to integrate task-specific knowledge into large vision backbones like DETR and Mask R-CNN [188, 218, 41]. This allows the base detection models to be repurposed for new object categories or specialized detection tasks (e.g., autonomous driving [772, 356, 452, 259, 58, 61], surveillance [643, 316, 156, 157, 120, 474]) without the need to retrain all parameters.

PEFT has also shown promise in vision-language tasks such as image captioning [300, 763, 238], visual question answering (VQA) [289, 20, 718, 192], and referring expression comprehension, where it helps adapt multimodal models like BLIP [366, 360, 57], Flamingo [11, 92, 453], and CLIP [121, 206, 585, 605] to specific domains or tasks. In such multimodal setups, PEFT modules can be injected into either the visual encoder, the language decoder, or their cross-attention layers to steer the joint representation learning efficiently.

Furthermore, PEFT facilitates continual learning in vision [734], enabling models to incorporate new classes or tasks incrementally without catastrophic forgetting. In few-shot and zero-shot image classification scenarios, PEFT makes it feasible to quickly adapt models with very limited supervision [310, 417, 235, 234, 498, 706].

As an example, PEFT techniques have enabled the integration of frozen visual encoders, such as SigLIP [781, 637], with pretrained LLMs for language-guided visual reasoning tasks. Similar to the X2L framework shown in **Figure 16**, the key innovation lies in the use of a lightweight two-layer multilayer perceptron (MLP) and a projection matrix P_θ, trained to convert visual features into a token-compatible format that can be understood by the LLM. This visual adapter module performs the necessary transformation without altering the vision encoder or language model, significantly reducing the number of parameters requiring training. This parameter-efficient approach supports diverse input formats—including static images, video frames, and multi-image sequences—making it highly suitable for tasks like temporal visual reasoning, multi-image comparison, and descriptive captioning. By grounding image features into a language-aligned semantic space, this PEFT-driven architecture ensures generalizability across domains and tasks without necessitating re-training of foundational models.

We present a comprehensive summary of PEFT methods applied in vision tasks across various LLMs and datasets in Tables 5, 6, 7, 13, and 14.

7.3 PEFT in Multimodal Learning

PEFT has become increasingly important in multimodal learning, where models process and integrate information from multiple input modalities—such as vision [561, 650, 25, 710, 134, 174, 768], language [113, 711, 214, 377, 118], audio [83, 37, 171, 119, 290, 112, 692], and video [542, 594, 361, 679, 237, 431, 682]. Modern multimodal architectures, like CLIP [121, 206, 585, 605], Flamingo [11, 92, 453], BLIP [366, 360, 57], PaLI [84], and Video-LLaMA [784], typically consist of large pretrained encoders and decoders spanning both visual and textual domains. Fine-tuning these models entirely is computationally expensive and memory-intensive, especially when adapting to new modalities, tasks, or domains. PEFT addresses this challenge by introducing lightweight, task-specific modules—such as adapters, low-rank matrices, or reparameterized prompts—into selected parts of the multimodal pipeline, allowing efficient and scalable adaptation.

In vision-language tasks like image captioning [300, 763, 649, 754, 546], visual question answering (VQA) [192, 300, 763, 238], and cross-modal retrieval [805, 680, 664, 688, 702, 168, 284], PEFT modules are often injected into the cross-attention layers between vision and text components,

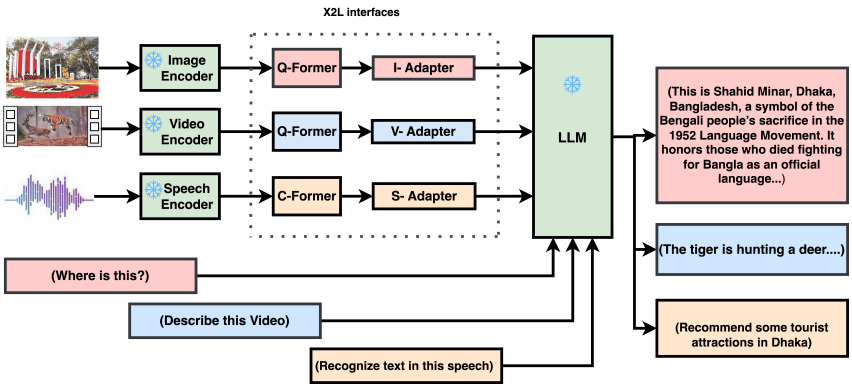


Figure 17: This figure outlines an X2L (Cross-modal to Language) framework where modality-specific encoders feed into lightweight adapters, which in turn align with a frozen LLM. Example outputs demonstrate cross-lingual generation grounded in images, videos, and speech, showcasing real-world applicability in multilingual and multimedia environments.

enabling the model to learn task-specific alignments without modifying the full backbone. Similarly, in video-language models used for tasks such as video QA [749, 779, 769, 95, 308, 349], temporal grounding [87, 383, 477, 350, 697, 798, 196], and action recognition [322, 270, 606, 676, 271, 231, 64], PEFT enables efficient fine-tuning on long video sequences by adapting only certain projection or fusion layers while freezing the majority of the vision encoder and language decoder.

Multimodal instruction tuning is another growing area where PEFT is heavily used, especially for aligning models to follow visual and language instructions together. In models like InstructBLIP [505, 690] and MiniGPT-4 [819, 755, 770, 23, 26], PEFT techniques allow fast customization to downstream multimodal tasks such as referring expression comprehension, image editing via text commands [54, 323, 162, 342, 572], and multimodal dialogue [285, 636, 379, 287, 659], all with limited supervision. Moreover, in low-resource or domain-specific settings (e.g., medical image–report generation or surveillance video QA), PEFT allows multimodal models to generalize effectively by training only a small subset of parameters.

As demonstrated in **Figure 17**, each modality—whether vision, audio, or text—is processed by a frozen encoder, such as SigLIP [781, 637] for images or pretrained audio models for speech. These encoders generate modality-specific embeddings, which are aligned into a shared latent space using transformer modules like Q-Former for visual features and C-Former for speech. Instead of re-training these components, PEFT introduces modality-specific lightweight adapters—namely, I-Adapter (image), V-Adapter (video), and S-Adapter (speech)—which serve as narrow bottleneck modules for mapping each modality's features into a unified token stream consumable by a frozen LLM. This approach localizes the adaptation to specific components, enabling the model to support cross-modal reasoning and generation tasks, such as bilingual captioning or voice-command interpretation, without degrading performance on prior capabilities. The decoupled design makes it possible to incrementally expand the system to new modalities with minimal overhead, exemplifying PEFT's utility in extensible and memory-efficient architectures.

7.4 PEFT in Robotics

In robotics, PEFT is increasingly being used to adapt large pretrained models—particularly vision-language and policy models—for control [433, 573, 803], perception [730, 499, 635, 483, 745, 725, 565], and decision-making tasks [297, 8, 473, 638, 640, 468] in physical environments. Robotics systems often require integrating visual input, natural language instructions, and low-level control signals to perform complex tasks in real-world settings. However, fine-tuning large multimodal models for each specific robot, environment, or task is often impractical due to computational and data constraints. PEFT provides a practical solution by allowing targeted fine-tuning of small subsets of parameters, such as adapters or low-rank projections, while keeping the majority of the base model frozen.

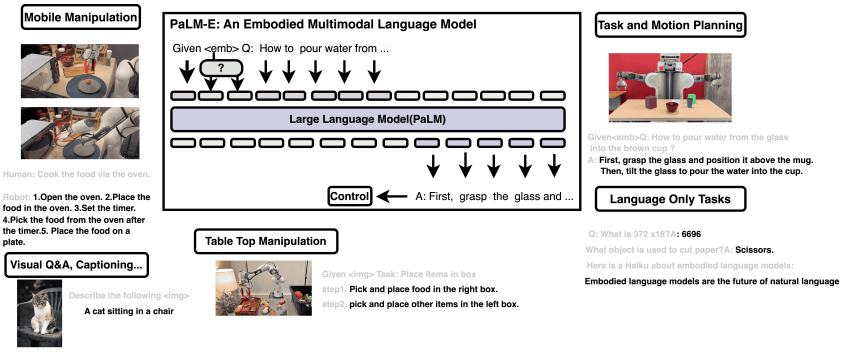


Figure 18: The figure presents various real-world robotic use cases, such as cooking, object sorting, and visual Q & A, all mediated through a unified LLM that integrates perception and control. By interpreting complex queries and translating them into actionable steps, the model exemplifies the future of grounded language understanding in robotics.

For instance, in vision-language-action models used in imitation learning and instruction following, PEFT enables adaptation to new environments or unseen tasks with minimal retraining. Techniques like LoRA [247, 133, 608, 49] or adapter tuning [421, 222, 789, 344] have been successfully applied in models such as RT-2 [53, 823] and SayCan [9], where robots are guided by high-level language commands grounded in visual context. PEFT also facilitates domain adaptation—e.g., transferring policies trained in simulation to the real world (sim-to-real)—by fine-tuning lightweight modules on real-world data without the need to retrain the entire policy network [158, 147, 24, 521, 714, 317, 550].

In embodied AI, where agents interact with their surroundings via sensors and actuators (e.g., navigation, manipulation, object fetching), PEFT allows for task- or goal-specific adaptation by introducing small trainable components into large pretrained transformers or diffusion policies [243, 201, 126, 548, 549]. These approaches help maintain generalization across environments while allowing fast adaptation to new robotic skills with limited data. Moreover, in multi-robot or multi-task scenarios, PEFT promotes modularity and parameter sharing, enabling efficient scalability across hardware platforms and task sets.

As illustrated in **Figure 18**, although current implementations like PaLM-E [149] are not explicitly PEFT-based, their architecture—a frozen LLM processing tokenized streams of language prompts, visual inputs, and proprioceptive data—lends itself well to PEFT augmentation. The figure demonstrates various real-world robotic applications including mobile manipulation, tabletop tasks, and motion planning, all unified through a central language model. In such systems, PEFT can be applied through LoRA (Low-Rank Adaptation) modules designed for specific output heads, or by introducing lightweight adapters that facilitate fusion of proprioceptive information. These adaptations would allow the robot to learn task-specific motor commands, such as "grasp," "rotate," or "navigate," with high efficiency and without retraining the entire network. This is particularly advantageous in robotic environments, where data is sparse and tasks are dynamic, requiring continual learning without compromising previously acquired behaviors. Through the integration of PEFT modules, robotics systems can achieve lifelong learning capabilities, extending their utility across diverse operational contexts with minimal retraining cost.

8 Complexity of PEFT Methods

Table 3 provides a detailed comparison of various PEFT methods based on their space and time complexity, as well as the total number of trainable parameters (TTPs) and additional parameters (APs) introduced during fine-tuning. Traditional full fine-tuning (FT) modifies all parameters of the model, resulting in a quadratic complexity of $O(d \times d)$ in both space and time, with a high memory footprint and zero additional modularity.

Adapter-based methods such as $(IA)^3$ reduce the fine-tuning burden by introducing small modules within the transformer layers, yielding linear complexity and maintaining trainable parameter counts at $3d$. Soft prompt-based methods, like Prompt and Prefix tuning, encode task-specific knowledge

Category	Method	Space Complexity	Time Complexity	TTPs	APs
Full Fine-Tuning	FT (Full Fine-Tuning)	$O(d \times d)$	$O(d \times d)$	d^2	0
Adapter-Based Fine-Tuning	$(IA)^3$	$O(l_k + l_v + l_{ff})$	$O(d_k + d_v + d_{ff})$	$3d$	$3d$
Soft Prompt-Based	Prompt	$O(d \times l_p)$	$O(d \times l_p)$	$l_p d$	$l_p d$
Soft Prompt-Based	Prefix	$O(L \times d \times l_p)$	$O(L \times d \times l_p)$	$Ll_p d$	$Ll_p d$
Structured Fine-Tuning	LoRA	$O((d+d) \times r)$	$O((d+d) \times r)$	$2dr$	dr
Structured Fine-Tuning	LoRA-FA	$O((d+d) \times r)$	$O((d+d) \times r)$	dr	$2dr$
Adaptive Rank Methods	AdaLoRA	$O((d+d+r) \times r)$	$O((d+d+r) \times r)$	$2dr + r^2$	$2dr + r^2$
Hybrid Approach	LoHA	$O(2r \times (d+d))$	$O(2r \times (d+d))$	$4dr$	$4dr$
Low Rank Decomposition	RoCoFT (Row)	$O(d \times r)$	$O(d \times r)$	rd	0
Low Rank Decomposition	RoCoFT (Column)	$O(d \times r)$	$O(d \times r)$	rd	0
Scaling Adaptation	Propulsion	$O(d)$	$O(d)$	d	d

Table 3: Comparison of PEFT methods and their computational complexity. Here, **TTPs** refers to the *total trainable parameters*, and **APs** refers to the *additional parameters* introduced by the fine-tuning method.

Method	ΔW Reparameterization	Notes
Intrinsic SAID	$\Delta W = F(W^r)$	F : $\mathbb{R}^r \to \mathbb{R}^d$, $W^r \in \mathbb{R}^r$ are parameters to be optimized, and $r \ll d$.
LoRA	$\Delta W = W_{\text{down}} W_{\text{up}}$	$W_{\text{down}} \in \mathbb{R}^{d \times r}$, $W_{\text{up}} \in \mathbb{R}^{r \times d}$, and $r \ll \{k, d\}$.
KronA	$\Delta W = W_{\text{down}} \otimes W_{\text{up}}$	rank($W_{\text{down}} \otimes W_{\text{up}}$) = rank($W_{\text{down}}$) $\times$ rank(W_{up}).
DyLoRA	$\Delta W = W_{\text{down}\downarrow b} W_{\text{up}\downarrow b}$	$W_{\text{down}\downarrow b} = W_{\text{down}}[: \; b, :]$, $W_{\text{up}\downarrow b} = W_{\text{up}}[:, : \; b]$, $b \in \{r_{\text{min}}, \ldots, r_{\text{max}}\}$.
AdaLoRA	$\Delta W = PAQ$	$PP^\top = P^\top P \neq I = QQ^\top = Q^\top Q$, $\Lambda = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$.
IncreLoRA	$\Delta W = W_{\text{down}} \Lambda W_{\text{up}}$	$\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_r]$ with λ_i being an arbitrary constant.
DeltaLoRA	$\Delta W = W_{\text{down}} W_{\text{up}}$	$W^{(t+1)} \leftarrow W^{(t)} + \left(W_{\text{down}}^{(t+1)} W_{\text{up}}^{(t+1)} - W_{\text{down}}^{(t)} W_{\text{up}}^{(t)}\right)$.
LoRAPrune	$\Delta W = W_{\text{down}} W_{\text{up}} \odot M$	$\delta = (W + W_{\text{down}} W_{\text{up}}) \odot M$, $M \in \{0,1\}^{1 \times G}$, G is group number.
QLoRA	$\Delta W = W_{\text{down}}^{BF16} W_{\text{up}}^{BF16}$	$Y^{BF16} = X^{BF16} \cdot \text{doubleDequant}(c_1^{FP32}, c_2^{FP8}, W^{NF4}) + X^{BF16} W_{\text{down}}^{BF16} W_{\text{up}}^{BF16}$.
QA-LoRA	$\Delta W = W_{\text{down}} W_{\text{up}}$	$W_{\text{down}} \in \mathbb{R}^{d \times r}$, $W_{\text{up}} \in \mathbb{R}^{r \times L}$, L is the quantization group number of W.
LoFTQ	$\Delta W = SVD(W - Q_t)$	$Q_t = q_N\left(W - W_{\text{down}}^{t-1} W_{\text{up}}^{t-1}\right)$, q_N is N-bit quantization function.
Kernel-mix	$\Delta W^h = \left(B_{\text{LoRA}}, B^h\right) \begin{pmatrix} A^h_{\text{LoRA}} \\ A^h \end{pmatrix}$	B_{LoRA} is shared across all heads, B^h, A^h provide rank-r update in each head..
LoRA-FA	$\Delta W = W_{\text{down}} W_{\text{up}} = QRW_{\text{up}}$	W_{down} is frozen, and only W_{up} is updated.
RoCoFT	$\Delta W = W_0 + R$ $\Delta W = W_0 + C$	R and C are restricted weight matrices such that only at most r of the rows or columns are non-zero.
Propulsion	$\Delta W = (WX) \odot Z$	W is frozen, X is input, and Z is propulsion parameters.

Table 4: Comparison of delta weight reparameterization across various PEFT methods. Representations of the baseline methods are taken from [739].

into learnable token embeddings, with complexity tied to the prompt length l_p and number of layers L. These methods allow for highly modular adaptation while keeping training costs manageable.

Structured fine-tuning approaches like LoRA and its variant LoRA-FA factorize weight updates into low-rank matrices, reducing the number of trainable parameters to $O(dr)$ or less. Adaptive rank methods, such as AdaLoRA, dynamically adjust the rank during training, offering a flexible trade-off between performance and efficiency, though at a slightly higher parameter count due to the inclusion of r^2 terms.

Hybrid approaches like LoHA further extend the expressiveness of LoRA by introducing hierarchical adaptation, doubling the parameter footprint ($4dr$) in exchange for better task generalization. Similarly, RoCoFT applies a low-rank decomposition at either the row or column level of weight matrices, maintaining very low complexity ($O(d \times r)$) with no additional overhead beyond trainable parameters.

Finally, Propulsion represents an extremely lightweight and scalable fine-tuning mechanism, introducing only $O(d)$ space and time complexity, with both TTPs and APs capped at d. This makes it particularly attractive for edge and low-resource deployment.

In addtion, Table 4 provides a comprehensive comparison of various PEFT methods based on their reparameterization of the delta weight matrix ΔW. Each method uses different strategies for

adjusting the weight updates during fine-tuning, optimizing parameter efficiency while maintaining performance.

Overall, the table illustrates the diverse trade-offs between efficiency, modularity, and expressivity across PEFT techniques, offering a toolkit of strategies tailored to specific deployment constraints and task complexities.

9 Strengths and Weaknesses of PEFT

PEFT has emerged as a transformative approach in adapting large pre-trained models to downstream tasks, offering a compelling balance between computational efficiency [325, 247, 788, 529, 215] and task-specific performance. One of its primary strengths lies in its ability to significantly reduce computational and memory costs by updating only a small subset of the model's parameters or introducing lightweight adapters, making it feasible to fine-tune large models on resource-constrained hardware [327, 250, 775]. This efficiency extends to faster training times and lower energy consumption, which is particularly advantageous in environmentally conscious applications. Additionally, PEFT mitigates the risk of catastrophic forgetting by preserving the general knowledge encoded in the pre-trained model, while still enabling effective transfer learning, especially in low-data regimes [327, 775].

However, PEFT is not without its limitations. It may underperform in tasks requiring significant adaptation of the pre-trained model, as the constraints imposed by limited parameter updates can restrict the model's ability to fully capture task-specific nuances [402, 384, 813]. Furthermore, some PEFT methods introduce architectural complexity, making implementation and debugging more challenging compared to standard fine-tuning. The approach can also be sensitive to hyperparameters, such as the size of adapter layers or the rank of low-rank approximations, necessitating extensive experimentation to achieve optimal performance [421, 808]. Additionally, PEFT may struggle with tasks that require a drastic shift from the pre-training Application, as it is most effective when the downstream task is closely related to the original training data [523].

Despite these challenges, PEFT remains a powerful tool for scaling large models across diverse applications, and ongoing research aims to address its limitations, such as improving flexibility for diverse tasks and reducing hyperparameter sensitivity, to further enhance its utility in the field of machine learning.

10 Discussion

Despite the remarkable progress of PEFT techniques in reducing computational and memory demands for adapting large language and vision models, several pressing challenges remain unresolved. Current methods often rely on heuristics rather than principled understanding, leading to inconsistent performance across tasks, architectures, and modalities. The lack of theoretical grounding regarding parameter sensitivity, the opaque nature of learned prompts and adapter modules, and the absence of unified benchmarks hinder reproducibility and generalization. Moreover, most PEFT approaches operate without consideration of task structure, domain knowledge, or semantic alignment—resulting in adaptations that, while efficient, are often suboptimal or cognitively naive. These limitations highlight the need for deeper analysis of model internals, architecture-aware design, and standardized evaluation to realize the full potential of PEFT in real-world, multimodal, and evolving scenarios.

11 Future Research Directions

PEFT methods have emerged as essential tools for adapting large-scale foundation models under computational and storage constraints, the current trajectory of research reveals several key areas where further investigation is both necessary and promising. These directions are outlined below to guide the evolution of PEFT toward greater generalizability, robustness, and theoretical maturity.

11.1 Theoretical Understanding of Parameter Influence

Most PEFT methods are grounded in empirical success rather than analytical rigor. Future research must prioritize the development of theoretical frameworks that explain how small subsets of trainable

parameters influence overall model adaptation. Concepts from information theory, such as mutual information between adapted modules and output prediction, or from optimization theory, such as curvature of the loss landscape around modular updates, could be leveraged to quantify adaptation efficiency. A better theoretical grounding would not only enhance interpretability but also inform principled design choices across diverse architectures.

11.2 Layer-wise Sensitivity and Structural Adaptation

In transformer-based architectures, not all layers contribute equally to downstream task performance. Existing PEFT approaches often insert adapter modules or low-rank projections uniformly across layers, which may be suboptimal. Future work should explore sensitivity-based placement strategies—using tools such as Jacobian analysis or Fisher Information Matrix estimates—to identify layers where fine-tuning yields the highest performance-to-parameter ratio. Additionally, research should focus on adaptive placement strategies, where modules are dynamically activated based on input complexity or layer activation statistics during training.

11.3 Task-Aware and Domain-Specific PEFT

While current PEFT methods are generally task-agnostic, real-world applications often involve domain-specific constraints and task structures. For example, tasks in legal or medical NLP involve complex semantic dependencies, while vision tasks in robotics may require temporally aligned fine-tuning. Future PEFT frameworks should incorporate inductive biases tailored to task semantics, perhaps through structure-aware prompts, hierarchical adapters, or task-conditioned reparameterization schemes. Integrating symbolic reasoning elements, causal graphs, or domain ontologies may also enhance generalization in low-data or high-stakes scenarios.

11.4 Generalization to Multimodal and Non-Transformer Architectures

Most PEFT techniques have been developed and tested primarily on large transformer-based LLMs. However, an increasing number of vision models (e.g., CNN-Transformer hybrids) and multimodal architectures (e.g., CLIP, Flamingo, Gato) demand adaptation strategies that account for modality entanglement, asynchronous inputs, and stream-wise attention fusion. Future research should design PEFT modules that maintain cross-modal coherence, minimize information bottlenecks, and support modality-specific adaptation while preserving inter-modal alignment. Exploration of fine-tuning strategies for non-transformer backbones, such as graph neural networks or recurrent models, also remains largely uncharted.

11.5 Continual and Lifelong Learning Integration

PEFT methods are typically designed for static, single-task adaptation. However, real-world environments demand continual adaptation to evolving tasks and distributions. Incorporating lifelong learning principles—such as replay-based memory modules, regularization-based knowledge retention, or dynamically growing parameter banks—into PEFT frameworks would enable more resilient and context-aware models. Sparse adapter stacking, delta compression, and orthogonal subspace training are promising avenues for enabling memory-efficient continual PEFT without catastrophic forgetting.

11.6 Interpretability and Explainability of PEFT Modules

The modular nature of PEFT methods presents an opportunity for improved interpretability, yet this potential remains underexploited. Few studies have systematically investigated what adapter layers or learned prompts actually encode. Future work should develop attribution techniques and visualization tools to trace the flow of information through PEFT modules. Interpretable tuning may involve aligning adapter activations with human-understandable concepts, analyzing prompt token behavior across tasks, or quantifying attention shifts induced by fine-tuning. Such developments are particularly crucial in applications where explainability is legally or ethically mandated.

11.7 Privacy-Preserving and Federated PEFT

The intersection of PEFT with differential privacy and federated learning is a promising but underdeveloped area. Given the proliferation of LLM deployment in privacy-sensitive contexts—such as healthcare, finance, and education—future research must explore methods to fine-tune models without centralized data access. Approaches like differentially private LoRA, secure adapter aggregation, or decentralized prompt tuning may offer viable paths forward. These methods should aim to maintain fine-tuning efficiency while rigorously protecting user data and ensuring compliance with regulatory standards.

11.8 Standardization of Benchmarks and Evaluation Protocols

There is an urgent need for standardized, multimodal benchmark suites designed specifically for evaluating PEFT methods. These should span diverse task types (e.g., classification, generation, reasoning), data regimes (low-resource, zero-shot, few-shot), and domains (general-purpose, biomedical, legal). Additionally, evaluation protocols should include robustness tests under domain shift, noise injection, and adversarial perturbations. Establishing such benchmarks will enhance reproducibility, allow fair comparisons, and accelerate the iterative improvement of PEFT methodologies.

11.9 Hardware-Aware and Sustainable PEFT

As AI systems are increasingly deployed on edge devices and in environmentally sensitive settings, PEFT research must align with the goals of hardware-awareness and energy efficiency. Techniques should be optimized for specific accelerators (e.g., TPUs, NPUs, FPGAs), and evaluated not only on accuracy and parameter count but also on latency, power consumption, and carbon footprint. Green AI practices, including low-bit quantized PEFT modules or adaptive update schedules that terminate early on easy samples, may contribute to more sustainable large-scale model use.

11.10 Meta-PEFT: Learning to Tune Efficiently

A promising meta-direction involves designing systems that automatically learn how to fine-tune models efficiently. Meta-PEFT approaches may employ reinforcement learning, neural architecture search, or gradient-based meta-learning to discover optimal PEFT strategies across tasks and models. This could lead to generalizable policies for adapter placement, prompt design, or rank selection, significantly reducing manual trial-and-error and improving portability across diverse domains.

12 Conclusion

As the scale and ubiquity of large language, vision, and multimodal models continue to expand, the demand for computationally efficient and scalable fine-tuning strategies has become increasingly urgent. PEFT techniques have emerged as a pragmatic and powerful response to these demands, enabling adaptation of large-scale models to diverse downstream tasks with minimal additional resource overhead. This survey has provided a comprehensive synthesis of PEFT methodologies, categorizing them into additive, selective, reparameterized, hybrid, and unified frameworks. By analyzing their design principles, parameter behaviors, and architectural integration, we have highlighted the core mechanisms that underlie their effectiveness.

We have also demonstrated the broad applicability of PEFT methods across language processing, visual understanding, and generative modeling, emphasizing how these strategies bridge the gap between performance and efficiency. Moreover, we have identified critical challenges in areas such as interpretability, task generalization, continual learning, and theoretical grounding. Addressing these challenges will be essential for building adaptive, robust, and sustainable AI systems.

Looking forward, the role of PEFT is poised to become even more central in future AI development—particularly in settings where privacy, environmental constraints, or domain specificity limit the feasibility of traditional fine-tuning. By distilling the current landscape and charting key research directions, this work aims to serve as a foundational reference for researchers and practitioners committed to advancing efficient, equitable, and accessible model adaptation in the era of foundation models.

References

[1] S Aathilakshmi, G Sivapriya, and T Manikandan. 6 llm fine-tuning: Instruction and parameter-efficient fine-tuning (peft). *Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks*, page 117, 2024.

[2] Laith Abualigah, Mohammad Qassem Bashabsheh, Hamzeh Alabool, and Mohammad Shehab. Text summarization: a brief review. *Recent Advances in NLP: the case of Arabic language*, pages 1–15, 2019.

[3] Laith Abualigah, Mohammad Qassem Bashabsheh, Hamzeh Alabool, and Mohammad Shehab. Text summarization: a brief review. *Recent Advances in NLP: the case of Arabic language*, pages 1–15, 2020.

[4] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Belarmino Adenso-Diaz and Manuel Laguna. Fine-tuning of algorithms using fractional experimental designs and local search. *Operations research*, 54(1):99–114, 2006.

[7] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[8] Alejandro Agostini, Carme Torras, and Florentin Wörgötter. Efficient interactive decision-making framework for robotic applications. *Artificial Intelligence*, 247:187–212, 2017.

[9] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[10] Rasmieh Al-Amer, Lucie Ramjan, Paul Glew, Maram Darwish, and Yenna Salamonson. Translation of interviews from a source language to a target language: Examining issues in cross-cultural health care research. *Journal of clinical nursing*, 24(9-10):1151–1162, 2015.

[11] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[12] Sergio Albeverio and Konstantin A Makarov. Nontrivial attractors in a model related to the three-body quantum problem. *Acta Applicandae Mathematica*, 48:113–184, 1997.

[13] Asmer Hamid Ali, Fan Zhang, Li Yang, and Deliang Fan. Learning to prune and low-rank adaptation for compact language model deployment. In *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, pages 36–42, 2025.

[14] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[15] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019.

[16] Yali Amit, Pedro Felzenszwalb, and Ross Girshick. Object detection. In *Computer vision: A reference guide*, pages 875–883. Springer, 2021.

[17] Roberto Omar Andrade and Sang Guun Yoo. A comprehensive study of the use of lora in the development of smart cities. *Applied Sciences*, 9(22):4753, 2019.

[18] Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. *arXiv preprint arXiv:2110.07560*, 2021.

[19] Alan Ansell, Ivan Vulić, Hannah Sterz, Anna Korhonen, and Edoardo M Ponti. Scaling sparse fine-tuning to large language models. *arXiv preprint arXiv:2401.16405*, 2024.

[20] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[21] Shrikara Arun, Meghana Tedla, and Karthik Vaidhyanathan. Llms for generation of architectural components: An exploratory empirical study in the serverless world. *arXiv preprint arXiv:2502.02539*, 2025.

[22] Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. Attempt: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. *arXiv preprint arXiv:2205.11961*, 2022.

[23] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

[24] Michael M Atkinson and William D Coleman. Policy networks, policy communities and the problems of governance. *Governance*, 5(2):154–180, 1992.

[25] Lloyd L Avant. Vision in the ganzfeld. *Psychological Bulletin*, 64(4):246, 1965.

[26] Vahid Azizi and Fatemeh Koochaki. Minigpt-reverse-designing: Predicting image adjustments utilizing minigpt-4. *arXiv preprint arXiv:2406.00971*, 2024.

[27] Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. Understanding multi-head attention in abstractive summarization. *arXiv preprint arXiv:1911.03898*, 2019.

[28] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[29] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[30] Dana Harry Ballard and Christopher M Brown. *Computer vision*. Prentice Hall Professional Technical Reference, 1982.

[31] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

[32] Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*, 2019.

[33] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2015.

[34] Samyadeep Basu, Shell Hu, Daniela Massiceti, and Soheil Feizi. Strong baselines for parameter-efficient few-shot fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11024–11031, 2024.

[35] Quentin Bateux, Jonathan Koss, Patrick W Sweeney, Erika Edwards, Nelson Rios, and Aaron M Dollar. Improving the accuracy of automated labeling of specimen images datasets via a confidence-based process. *arXiv preprint arXiv:2411.10074*, 2024.

[36] Emily M. Bender. 100 things you always wanted to know about linguistics but were afraid to ask*. In Radu Florian and Jacob Eisenstein, editors, *Tutorial Abstracts at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, June 2012. Association for Computational Linguistics.

[37] Christian Benoit, Jean-Claude Martin, Catherine Pelachaud, Lambert Schomaker, and Bernhard Suhm. Audio-visual and multimodal speech systems. *Handbook of Standards and Resources for Spoken Language Systems-Supplement*, 500:1–95, 2000.

[38] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.

[39] James Bernhard. Alternatives to the scaled dot product for attention in the transformer neural network architecture. *arXiv preprint arXiv:2311.09406*, 2023.

[40] Jacob Beutel. *Handbook of medical imaging*, volume 3. Spie Press, 2000.

[41] Puja Bharati and Ankita Pramanik. Deep learning techniques—r-cnn to mask r-cnn: a survey. *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pages 657–668, 2020.

[42] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pages 864–873. PMLR, 2020.

[43] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[44] Bohdan Bilonoh and Sergii Mashtalir. Parallel multi-head dot product attention for video summarization. In *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pages 158–162. IEEE, 2020.

[45] Prayma Bishshash, Asraful Sharker Nirob, Habibur Shikder, Afjal Hossan Sarower, Touhid Bhuiyan, and Sheak Rashed Haider Noori. A comprehensive cotton leaf disease dataset for enhanced detection and classification. *Data in Brief*, 57:110913, 2024.

[46] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

[47] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[48] Yelysei Bondarenko, Riccardo Del Chiaro, and Markus Nagel. Low-rank quantization-aware training for llms. *arXiv preprint arXiv:2406.06385*, 2024.

[49] Martin C Bor, John Vidler, and Utz Roedig. Lora for the internet of things. In *Ewsn*, volume 16, pages 361–366, 2016.

[50] Surekha Borra, Rohit Thanki, and Nilanjan Dey. *Satellite image analysis: clustering and classification*. Springer, 2019.

[51] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.

[52] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pages 1059–1071. PMLR, 2021.

[53] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[54] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[55] Anthony GA Brown, Antonella Vallenari, Timo Prusti, Jos HJ De Bruijne, Carine Babusiaux, Michael Biermann, Orlagh L Creevey, Dafydd Wyn Evans, Laurent Eyer, A Hutton, et al. Gaia early data release 3-summary of the contents and survey properties. *Astronomy & Astrophysics*, 649:A1, 2021.

[56] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[57] Miriam Cabero, Andrew Lundgren, Alex H Nitz, Thomas Dent, David Barker, Evan Goetz, Jeff S Kissel, Laura K Nuttall, Paul Schale, Robert Schofield, et al. Blip glitches in advanced ligo data. *Classical and Quantum Gravity*, 36(15):155010, 2019.

[58] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[59] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018.

[60] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, Antonio Feraco, et al. *A practical guide to sentiment analysis*, volume 5. Springer, 2017.

[61] Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4649–4672, 2010.

[62] Hu Cao, Zhongnan Qu, Guang Chen, Xinyi Li, Lothar Thiele, and Alois Knoll. Ghostvit: Expediting vision transformers via cheap operations. *IEEE Transactions on Artificial Intelligence*, 5(6):2517–2525, 2023.

[63] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[64] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[65] Xavier Carreras and Lluis Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 152–164, 2005.

[66] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[67] Shurong Chai, Rahul Kumar Jain, Shiyu Teng, Jiaqing Liu, Yinhao Li, Tomoko Tateyama, and Yen-wei Chen. Ladder fine-tuning approach for sam integrating complementary network. *Procedia Computer Science*, 246:4951–4958, 2024.

[68] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[69] Amélie Chatelain, Amine Djeghri, Daniel Hesslow, and Julien Launay. Is the number of trainable parameters all that actually matters? In *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*, pages 27–32. PMLR, 2022.

[70] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE, 2018.

[71] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19133–19143, 2023.

[72] Dongping Chen. Aggregate, decompose, and fine-tune: A simple yet effective factor-tuning method for vision transformer. *arXiv preprint arXiv:2311.06749*, 2023.

[73] Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Xiang Li, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-adapter: Exploring parameter efficient transfer learning for convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1551–1561, 2024.

[74] Haonan Chen, Jiaming Xu, Lily Sheng, Tianchen Ji, Shuijing Liu, Yunzhu Li, and Katherine Driggs-Campbell. Learning coordinated bimanual manipulation policies using state diffusion and inverse dynamics models. *arXiv preprint arXiv:2503.23271*, 2025.

[75] Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:2301.01821*, 2023.

[76] Lei Chen, Houwei Chou, and Xiaodan Zhu. Developing prefix-tuning models for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 390–397, 2022.

[77] Leiyu Chen, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712, 2021.

[78] Lichang Chen, Heng Huang, and Minhao Cheng. Ptp: Boosting stability and performance of prompt tuning with perturbation-based regularizer. *arXiv preprint arXiv:2305.02423*, 2023.

[79] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

[80] Long Chen, Oreste Villa, Sriram Krishnamoorthy, and Guang R Gao. Dynamic load balancing on single-and multi-gpu systems. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*, pages 1–12. IEEE, 2010.

[81] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[82] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[83] Tsuhan Chen and Ram R Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[84] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[85] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552*, 2018.

[86] Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. Tigerbot: An open multilingual multitask llm. *arXiv preprint arXiv:2312.08688*, 2023.

[87] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453, 2021.

[88] Yuliang Chen. *Efficient Mixed-Precision Quantization of Deep Neural Networks for Edge Applications*. PhD thesis, Politecnico di Torino, 2024.

[89] Yuyan Chen, Qiang Fu, Ge Fan, Lun Du, Jian-Guang Lou, Shi Han, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hadamard adapter: An extreme parameter-efficient adapter tuning method for pre-trained language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 276–285, 2023.

[90] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[91] Gong Cheng, Junwei Han, and Lei Guo. Remote sensing image scene classification: Benchmark and state of the art. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[92] B Childress, F Arengo, A Bechet, and N Jarrett. Flamingo. *Bulletin of the IUCN-SSC/Wetlands International Flamingo Specialist Group*, 17, 2008.

[93] Rajas Chitale, Ankit Vaidya, Aditya Kane, and Archana Ghotkar. Task arithmetic with lora for continual learning. *arXiv preprint arXiv:2311.02428*, 2023.

[94] Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok, and SangKeun Lee. Smop: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14306–14316, 2023.

[95] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 1166–1174, 2021.

[96] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[97] Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR, 2023.

[98] Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*, 2023.

[99] Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.

[100] Yao-Shun Chuang, Xiaoqian Jiang, Chun-Teh Lee, Ryan Brandon, Duong Tran, Oluwabunmi Tokede, and Muhammad F Walji. Use gpt-j prompt generation with roberta for ner models on diagnosis extraction of periodontal diagnosis from electronic dental records. In *AMIA Annual Symposium Proceedings*, volume 2023, page 904, 2024.

[101] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[102] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

[103] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

[104] K Clark. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[105] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[106] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[107] Richard E Clark and Alexander Voogel. Transfer of training principles for instructional design. *Ectj*, 33(2):113–123, 1985.

[108] Christopher Clarke, Yuzhao Heng, Lingjia Tang, and Jason Mars. Peft-u: Parameter-efficient fine-tuning for user personalization. *arXiv preprint arXiv:2407.18078*, 2024.

[109] Adam Coates, Honglak Lee, and Andrew Y Ng. An analysis of single-layer networks in unsupervised feature learning. Technical report, Stanford University, 2011. STL-10 Dataset Technical Report.

[110] Adam Coates, Honglak Lee, and Andrew Y. Ng. Stl-10: A dataset for developing more effective unsupervised learning methods. Technical report, Stanford University, 2011.

[111] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[112] Maximo Cobos, Jens Ahrens, Konrad Kowalczyk, and Archontis Politis. An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):10, 2022.

[113] William W Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems (TOIS)*, 18(3):288–321, 2000.

[114] Bai Cong, Nico Daheim, Yuesong Shen, Daniel Cremers, Rio Yokota, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational low-rank adaptation using ivon. *arXiv preprint arXiv:2411.04421*, 2024.

[115] A Conneau. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[116] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[117] Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2:1–24, 2015.

[118] Linda M Crawford-Lange and Dale L Lange. Integrating language and culture: How to do it. *Theory into Practice*, 26(4):258–266, 1987.

[119] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4):1–46, 2016.

[120] Rita Cucchiara. Multimedia surveillance systems. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 3–10, 2005.

[121] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

[122] Nadia Daneshvar, Deepti Pandita, Shari Erickson, Lois Snyder Sulmasy, Matthew DeCamp, ACP Medical Informatics Committee, Professionalism the Ethics, and Human Rights Committee. Artificial intelligence in the provision of health care: an american college of physicians policy position paper. *Annals of Internal Medicine*, 177(7):964–967, 2024.

[123] Paul M Dare. Shadow analysis in high-resolution satellite imagery of urban areas. *Photogrammetric Engineering & Remote Sensing*, 71(2):169–177, 2005.

[124] Arijit Das. Natural galore: Accelerating galore for memory-efficient llm training and finetuning. *arXiv preprint arXiv:2410.16029*, 2024.

[125] Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, Peng Shi, Wenpeng Yin, and Rui Zhang. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. *arXiv preprint arXiv:2311.03748*, 2023.

[126] Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*, 2024.

[127] F David. A model of random surfaces with non-trivial critical behaviour. *Nuclear Physics B*, 257:543–576, 1985.

[128] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515, 2017.

[129] Gustavo H De Rosa and Joao P Papa. A survey on text generation using generative adversarial networks. *Pattern Recognition*, 119:108098, 2021.

[130] Silvia Demetri, Marco Zúñiga, Gian Pietro Picco, Fernando Kuipers, Lorenzo Bruzzone, and Thomas Telkamp. Automated estimation of link quality for lora: A remote sensing approach. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, pages 145–156, 2019.

[131] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[132] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022.

[133] Shilpa Devalal and A Karthikeyan. Lora technology-an overview. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pages 284–290. IEEE, 2018.

[134] Russell L DeValois and Karen K DeValois. *Spatial Vision*. Oxford University Press, 1990.

[135] M Devi Devika, C[a] Sunitha, and Amal Ganesh. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 87:44–49, 2016.

[136] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[137] Haiwen Diao, Bo Wan, Ying Zhang, Xu Jia, Huchuan Lu, and Long Chen. Unipt: Universal parallel tuning for transfer learning with efficient parameter and memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28729–28740, 2024.

[138] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents, 2019.

[139] Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. *arXiv preprint arXiv:2311.11696*, 2023.

[140] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[141] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. *arXiv preprint arXiv:2210.01033*, 2022.

[142] Haoyu Dong and Jianhong Shun. Low-rank adaptation for scalable fine-tuning of pre-trained language models. *preprints*, 2025.

[143] Ming Dong, Kang Xue, Bolong Zheng, and Tingting He. Data-oriented dynamic fine-tuning parameter selection strategy for fish mask based efficient fine-tuning. *arXiv preprint arXiv:2403.08484*, 2024.

[144] Ming Dong, Kang Xue, Bolong Zheng, and Tingting He. Targeted efficient fine-tuning: Optimizing parameter updates with data-driven sample selection. *arXiv preprint arXiv:2403.08484*, 2024.

[145] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.

[146] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[147] Keith Dowding. Model or metaphor? a critical review of the policy network approach. *Political studies*, 43(1):136–158, 1995.

[148] Caitlin Dreisbach, Theresa A Koleck, Philip E Bourne, and Suzanne Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125:37–46, 2019.

[149] Danny Driess, Fei Xia, Mehdi S.M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[150] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[151] Jiangfei Duan, Shuo Zhang, Zerui Wang, Lijuan Jiang, Wenwen Qu, Qinghao Hu, Guoteng Wang, Qizhen Weng, Hang Yan, Xingcheng Zhang, et al. Efficient training of large language models on distributed infrastructures: A survey. *arXiv preprint arXiv:2407.20018*, 2024.

[152] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[153] Gaute Klakegg Dvergsdal. Long-context llms: How extended context length is utilized, and its potential for refining decompiled code. Master's thesis, NTNU, 2024.

[154] Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza-Ruiz, et al. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*, 2021.

[155] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021.

[156] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77:103116, 2021.

[157] T Grace Emori, David H Culver, Teresa C Horan, William R Jarvis, John W White, David R Olson, Shailen Banerjee, Jonathan R Edwards, William J Martone, Robert P Gaynes, et al. National nosocomial infections surveillance system (nnis): description of surveillance methods. *American journal of infection control*, 19(1):19–35, 1991.

[158] Henrik Enroth. Policy network theory. *The SAGE handbook of governance*, pages 19–35, 2011.

[159] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[160] Nicolaas (KLAAS) M Faber. A closer look at the bias–variance trade-off in multivariate calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(2):185–192, 1999.

[161] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

[162] Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. Text editing by command. *arXiv preprint arXiv:2010.12826*, 2020.

[163] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2777–2787. IEEE, 2020.

[164] Lishui Fan, Jiakun Liu, Zhongxin Liu, David Lo, Xin Xia, and Shanping Li. Exploring the capabilities of llms for code change related tasks. *arXiv preprint arXiv:2407.02824*, 2024.

[165] Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. Automated federated pipeline for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2404.06448*, 2024.

[166] Farzad Farhadzadeh, Debasmit Das, Shubhankar Borse, and Fatih Porikli. Lora-x: Bridging foundation models with training-free cross-model adaptation. *arXiv preprint arXiv:2501.16559*, 2025.

[167] Li Fei-Fei, Marco Andreetto, and Marc'Aurelio Ranzato. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[168] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.

[169] Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. Mixture-of-loras: An efficient multitask tuning for large language models. *arXiv preprint arXiv:2403.03432*, 2024.

[170] Yue Feng and Yan Cheng. Short text sentiment analysis based on multi-channel cnn with multi-head attention mechanism. *IEEE Access*, 9:19854–19863, 2021.

[171] Jonathan Foote. An overview of audio information retrieval. *Multimedia systems*, 7(1):2–10, 1999.

[172] Common Crawl Foundation. Common crawl corpus. https://commoncrawl.org, 2008. Raw web data (HTML, WARC/WET format), 386 TiB.

[173] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*, 2022.

[174] Roger Fry. *Vision and design*. Courier Corporation, 2012.

[175] Chin-Lun Fu, Zih-Ching Chen, Yun-Ru Lee, and Hung-yi Lee. Adapterbias: Parameter-efficient token-dependent representation shift for adapters in nlp tasks. *arXiv preprint arXiv:2205.00305*, 2022.

[176] Minghao Fu, Ke Zhu, and Jianxin Wu. Dtl: Disentangled transfer learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12082–12090, 2024.

[177] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12799–12807, 2023.

[178] Deva Kumar Gajulamandyam, Sainath Veerla, Yasaman Emami, Kichang Lee, Yuanting Li, Jinthy Swetha Mamillapalli, and Simon Shim. Domain specific finetuning of llms using peft techniques. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00484–00490. IEEE, 2025.

[179] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[180] Debashis Ganguly, Srabonti Chakraborty, Maricel Balitanas, and Tai-hoon Kim. Medical imaging: A review. In *International Conference on Security-Enriched Urban Computing and Smart Grid*, pages 504–516. Springer, 2010.

[181] Chao Gao and Sai Qian Zhang. Dlora: Distributed parameter-efficient fine-tuning solution for large language model. *arXiv preprint arXiv:2404.05182*, 2024.

[182] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[183] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11483–11493, 2023.

[184] Xinjian Gao, Zhao Zhang, Tingting Mu, Xudong Zhang, Chaoran Cui, and Meng Wang. Self-attention driven adversarial similarity learning network. *Pattern Recognition*, 105:107331, 2020.

[185] Ali Reza Ghasemi and Javad Salimi Sartakhti. Multilingual language models in persian nlp tasks: A performance comparison of fine-tuning techniques. *Journal of AI and Data Mining*, 13(1):107–117, 2025.

[186] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. *arXiv preprint arXiv:2104.08771*, 2021.

[187] Michail Giannakos and Mutlu Cukurova. The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology*, 54(5):1246–1267, 2023.

[188] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[189] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[190] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[191] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[192] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[193] Eleonora Grassucci, Aston Zhang, and Danilo Comminiello. Phnns: Lightweight neural networks via parameterized hypercomplex convolutions. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[194] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[195] Naibin Gu, Peng Fu, Xiyu Liu, Bowen Shen, Zheng Lin, and Weiping Wang. Light-peft: Lightening parameter-efficient fine-tuning via early pruning. *arXiv preprint arXiv:2406.03792*, 2024.

[196] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024.

[197] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2):73–84, 1998.

[198] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

[199] Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*, 2020.

[200] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5436–5447, 2022.

[201] Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim GJ Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control. *Advances in Neural Information Processing Systems*, 37:74182–74210, 2024.

[202] Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. Amazonqa: A review-based question answering task, 2019.

[203] Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. Amazonqa: A review-based question answering task. *arXiv preprint arXiv:1908.04364*, 2019.

[204] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[205] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.

[206] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):20, 2021.

[207] Abid Haleem, Mohd Javaid, and Ibrahim Haleem Khan. Current status and applications of artificial intelligence (ai) in medical field: An overview. *Current Medicine Research and Practice*, 9(6):231–237, 2019.

[208] John T Halloran, Manbir Gulati, and Paul F Roysdon. Mamba state-space models can be strong downstream learners. *arXiv preprint arXiv:2406.00209*, 2024.

[209] Jiayi Han, Liang Du, Hongwei Du, Xiangguo Zhou, Yiwen Wu, Weibo Zheng, and Donghong Han. Slim: Let llm learn more and forget less with soft lora and identity mixture. *arXiv preprint arXiv:2410.07739*, 2024.

[210] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192, 2022.

[211] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[212] Jitai Hao, WeiWei Sun, Xin Xin, Qi Meng, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. Meft: Memory-efficient fine-tuning through sparse adapter. *arXiv preprint arXiv:2406.04984*, 2024.

[213] Rui Hao, Linmei Hu, Weijian Qi, Qingliu Wu, Yirui Zhang, and Liqiang Nie. Chatllm network: More brains, more intelligence. *AI Open*, 2025.

[214] Robert J Hartsuiker and Martin J Pickering. Language integration in bilingual sentence production. *Acta psychologica*, 128(3):479–489, 2008.

[215] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

[216] Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11825–11835, 2023.

[217] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

[218] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[219] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[220] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. corr abs/2111.09543 (2021). *arXiv preprint arXiv:2111.09543*, 2021.

[221] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[222] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021.

[223] Shwai He, Liang Ding, Daize Dong, Miao Zhang, and Dacheng Tao. Sparseadapter: An easy approach for improving the parameter-efficiency of adapters. *arXiv preprint arXiv:2210.04284*, 2022.

[224] Shwai He, Run-Ze Fan, Liang Ding, Li Shen, Tianyi Zhou, and Dacheng Tao. Mera: Merging pretrained adapters for few-shot learning. *arXiv preprint arXiv:2308.15982*, 2023.

[225] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 817–825, 2023.

[226] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE JSTARS*, 12:2217–2226, 2019.

[227] James Henderson, Sebastian Ruder, et al. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, 2021.

[228] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[229] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[230] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[231] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[232] Jose-Luis Hervas-Oliver, Gregorio Gonzalez, Pedro Caja, and Francisca Sempere-Ripoll. Clusters and industrial districts: Where is the literature going? identifying emerging sub-fields of research. *European Planning Studies*, 23(9):1827–1872, 2015.

[233] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[234] Richard A Hogan. Issues and approaches in supervision. *Psychotherapy: Theory, Research & Practice*, 1(3):139, 1964.

[235] Elizabeth L Holloway and Susan Allstetter Neufeldt. Supervision: Its contributions to treatment efficacy. *Journal of consulting and clinical psychology*, 63(2):207, 1995.

[236] John J Hopfield. Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the national academy of sciences*, 84(23):8429–8433, 1987.

[237] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

[238] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

[239] Soneya Binta Hossain, Nan Jiang, Qiang Zhou, Xiaopeng Li, Wen-Hao Chiang, Yingjun Lyu, Hoan Nguyen, and Omer Tripp. A deep dive into large language models for automated bug localization and repair. *Proceedings of the ACM on Software Engineering*, 1(FSE):1471–1493, 2024.

[240] Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 523–533, 2014.

[241] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 541–549, 2017.

[242] Yihao Hou, Christoph Bert, Ahmed Gomaa, Godehard Lahmer, Daniel Hoefler, Thomas Weissmann, Raphaela Voigt, Philipp Schubert, Charlotte Schmitter, Alina Depardon, et al. Fine-tuning a local llama-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *arXiv preprint arXiv:2408.10715*, 2024.

[243] Zhi Hou, Tianyi Zhang, Yuwen Xiong, Hengjun Pu, Chengyang Zhao, Ronglei Tong, Yu Qiao, Jifeng Dai, and Yuntao Chen. Diffusion transformer policy. *arXiv preprint arXiv:2410.15959*, 2024.

[244] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[245] Godfrey N Hounsfield. Computed medical imaging. *Science*, 210(4465):22–28, 1980.

[246] Eduard Hovy. Text summarization. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 583–598. Oxford University Press, 2005.

[247] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[248] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[249] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12511–12518, 2024.

[250] Jiang Hu and Quanzheng Li. Adafish: Fast low-rank parameter-efficient fine-tuning by using second-order information. *arXiv preprint arXiv:2403.13128*, 2024.

[251] Shengding Hu, Zhen Zhang, Ning Ding, Yadao Wang, Yasheng Wang, Zhiyuan Liu, and Maosong Sun. Sparse structure search for parameter-efficient tuning. *arXiv preprint arXiv:2206.07382*, 2022.

[252] Shishuai Hu, Zehui Liao, and Yong Xia. Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. *arXiv preprint arXiv:2211.11514*, 2022.

[253] Yue Hu, Xinan Ye, Yifei Liu, Souvik Kundu, Gourav Datta, Srikar Mutnuri, Namo Asavisanu, Nora Ayanian, Konstantinos Psounis, and Peter Beerel. Firefly: A synthetic dataset for ember detection in wildfire. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3765–3769, 2023.

[254] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*, 2023.

[255] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

[256] Feihu Huang, Min Jiang, Fang Liu, Dian Xu, Zimeng Fan, and Yonghao Wang. Classification of heads in multi-head attention mechanisms. In *International Conference on Knowledge Science, Engineering and Management*, pages 681–692. Springer, 2022.

[257] Tengjun Huang. Efficient remote sensing with harmonized transfer learning and modality alignment. *arXiv preprint arXiv:2404.18253*, 2024.

[258] Wei-Cheng Huang, Chun-Fu Chen, and Hsiang Hsu. Ovor: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. *arXiv preprint arXiv:2402.04129*, 2024.

[259] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018.

[260] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.

[261] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

[262] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338, 2018.

[263] W John Hutchins. Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445. Elsevier, 1995.

[264] Mirjana Ivanovic and Miloš Radovanovic. Modern machine learning techniques and their applications. In *International conference on electronics, communications and networks*, 2015.

[265] Abhinav Jain, Swarat Chaudhuri, Thomas Reps, and Chris Jermaine. Prompt tuning strikes back: Customizing foundation models with low-rank prompt adaptation. *arXiv preprint arXiv:2405.15282*, 2024.

[266] Navya Jain, Zekun Wu, CRISTIAN ENRIQUE MUNOZ VILLALOBOS, Airlie Hilliard, Adriano Koshiyama, Emre Kazim, and Philip Colin Treleaven. From text to emoji: How peft-driven personality manipulation unleashes the emoji potential in llms. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

[267] Elenaz Janfaza, A Assemi, and SS Dehghan. Language, translation, and culture. In *International conference on language, medias and culture*, volume 33, pages 83–87, 2012.

[268] Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[269] Mohamed Salah Jebali, Anna Valanzano, Malathi Murugesan, Giacomo Veneri, and Giovanni De Magistris. Leveraging the regularizing effect of mixing industrial and open source data to prevent overfitting of llm fine tuning. In *International Joint Conference on Artificial Intelligence 2024 Workshop on AI Governance: Alignment, Morality, and Law*, 2024.

[270] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[271] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[272] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[273] Ni Jia, Tong Liu, Jiadi Chen, Ying Zhang, and Song Han. Task-agnostic adaptive activation scaling network for llms. *IEEE Access*, 2025.

[274] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[275] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Yiliang Lv, Deli Zhao, and Jingren Zhou. Rethinking efficient tuning methods from a unified perspective. *arXiv preprint arXiv:2303.00690*, 2023.

[276] Ziyu Jiang, Tianlong Chen, Xuxi Chen, Yu Cheng, Luowei Zhou, Lu Yuan, Ahmed Awadallah, and Zhangyang Wang. Dna: Improving few-shot transfer learning with low-rank decomposition and alignment. In *European Conference on Computer Vision*, pages 239–256. Springer, 2022.

[277] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[278] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1060–1068, 2023.

[279] Shibo Jie, Zhi-Hong Deng, Shixuan Chen, and Zhijuan Jin. Convolutional bypasses are better vision transformer adapters. In *ECAI 2024*, pages 202–209. IOS Press, 2024.

[280] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position paper: What can large language models tell us about time series analysis. *arXiv e-prints*, pages arXiv–2402, 2024.

[281] Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. Prollm: protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *bioRxiv*, pages 2024–04, 2024.

[282] Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27144–27153, 2024.

[283] Nitin Jindal and Bing Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190, 2007.

[284] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021.

[285] Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 376–383, 2002.

[286] Aravind K Joshi. Natural language processing. *Science*, 253(5025):1242–1249, 1991.

[287] Danica Jovanovic and Theo Van Leeuwen. Multimodal dialogue on social media. *Social Semiotics*, 28(5):683–699, 2018.

[288] Euna Jung, Jaekeol Choi, and Wonjong Rhee. Semi-siamese bi-encoder neural ranking model using lightweight fine-tuning. In *Proceedings of the ACM Web Conference 2022*, pages 502–511, 2022.

[289] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017.

[290] Navin Kamuni, Sathishkumar Chintala, Naveen Kunchakuri, Jyothi Swaroop Arlagadda Narasimharaju, and Venkat Kumar. Advancing audio fingerprinting accuracy with ai and ml: Addressing background noise and distortion challenges. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 341–345. IEEE, 2024.

[291] Fahdi Kanavati and Masayuki Tsuneki. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning. In *Medical Imaging with Deep Learning*, pages 338–353. PMLR, 2021.

[292] Andreas Kanavos, Christos Makris, and Evangelos Theodoridis. Topic categorization of biomedical abstracts. *International Journal on Artificial Intelligence Tools*, 24(01):1540004, 2015.

[293] Chien-Hao Kao, Chih-Chieh Chen, and Yu-Tza Tsai. Model of multi-turn dialogue in emotional chatbot. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 1–5. IEEE, 2019.

[294] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[295] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

[296] Hany Kasban, MAM El-Bendary, DH Salama, et al. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, 4(2):37–58, 2015.

[297] Tobias Kaupp, Alexei Makarenko, and Hugh Durrant-Whyte. Human–robot communication for collaborative decision making—a probabilistic approach. *Robotics and Autonomous Systems*, 58(5):444–456, 2010.

[298] Kornraphop Kawintiranon and Lisa Singh. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735, 2021.

[299] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[300] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897, 2019.

[301] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021.

[302] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International conference on algorithmic learning theory*, pages 597–619. PMLR, 2023.

[303] Dorothy Kenny. Machine translation. In *The Routledge handbook of translation and philosophy*, pages 428–445. Routledge, 2018.

[304] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[305] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[306] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Stanford dogs dataset. http://vision.stanford.edu/aditya86/ImageNetDogs/, 2011.

[307] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[308] Khushboo Khurana and Umesh Deshpande. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey. *IEEE Access*, 9:43799–43823, 2021.

[309] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 829–839. Association for Computational Linguistics, 2019.

[310] Sue M Kilminster and Brian C Jolly. Effective supervision in clinical practice settings: a literature review. *Medical education*, 34(10):827–840, 2000.

[311] Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36:36187–36207, 2023.

[312] Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36, 2024.

[313] Sanghyeon Kim, Hyunmo Yang, Yunghyun Kim, Youngjoon Hong, and Eunbyung Park. Hydra: Multi-head low-rank adaptation for parameter efficient fine-tuning. *Neural Networks*, page 106414, 2024.

[314] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[315] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[316] Douglas N Klaucke, James W Buehler, Stephen B Thacker, R Gibson Parrish, and Frederick L Trowbridge. Guidelines for evaluating surveillance systems. Technical report, Centers for Disease Control and Prevention (CDC), 1988.

[317] E-H Klijn. Policy networks: an overview. *Managing complex networks: Strategies for the public sector*, pages 15–34, 1997.

[318] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating residual and normalization layers into analysis of masked language models. *arXiv preprint arXiv:2109.07152*, 2021.

[319] Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, 2022.

[320] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[321] Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597, 2015.

[322] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[323] Gary E Kopec and Steven C Bagley. Editing images of text. In *Proceedings of the International Conference on Electronic Publishing, Document Manipulation & Typography, Gaithersburg, Maryland*, pages 207–220, 1990.

[324] Mikhail V Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

[325] Md Kowsher, Tara Esmaeilbeig, Chun-Nam Yu, Mojtaba Soltanalian, and Niloofar Yousefi. Rocoft: Efficient finetuning of large language models with row-column updates. *arXiv preprint arXiv:2410.10075*, 2024.

[326] Md Kowsher, Ritesh Panditi, Nusrat Jahan Prottasha, Prakash Bhat, Anupam Kumar Bairagi, and Mohammad Shamsul Arefin. Token trails: Navigating contextual depths in conversational ai with chatllm. In *International Conference on Applications of Natural Language to Information Systems*, pages 56–67. Springer, 2024.

[327] Md Kowsher, Nusrat Jahan Prottasha, and Prakash Bhat. Propulsion: Steering llm with tiny fine-tuning. *arXiv preprint arXiv:2409.10927*, 2024.

[328] Md Kowsher, Nusrat Jahan Prottasha, and Chun-Nam Yu. Does self-attention need separate weights in transformers? *arXiv preprint arXiv:2412.00359*, 2024.

[329] Md Kowsher, Md Shohanur Islam Sobuj, Asif Mahmud, Nusrat Jahan Prottasha, and Prakash Bhat. L-tuning: Synchronized label tuning for prompt and prefix in llms. *arXiv preprint arXiv:2402.01643*, 2023.

[330] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561. IEEE, 2013.

[331] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009. Technical Report.

[332] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260, 2024.

[333] Shantanu Kumar and Shruti Singh. Fine-tuning llama 3 for sentiment analysis: Leveraging aws cloud for enhanced performance. *SN Computer Science*, 5(8):1–8, 2024.

[334] Vimal Kumar, Priyam Srivastava, Ashay Dwivedi, Ishan Budhiraja, Debjani Ghosh, Vikas Goyal, and Ruchika Arora. Large-language-models (llm)-based ai chatbots: Architecture, in-depth analysis and their performance evaluation. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 237–249. Springer, 2023.

[335] Jenny Kunz. Train more parameters but mind their placement: Insights into language adaptation with peft. *arXiv preprint arXiv:2412.12674*, 2024.

[336] Minoru Kurita. Tensor fields and their parallelism. *Nagoya mathematical journal*, 18:133–151, 1961.

[337] Alina Kuznetsova, Hassan Rom, Neil Alldrin, and et al. The open images dataset v6: A million-scale benchmark for object detection and visual relationship. *IJCV*, 2020.

[338] Namju Kwak and Taesup Kim. X-peft: extremely parameter-efficient fine-tuning for extreme multi-profile scenarios. *arXiv preprint arXiv:2401.16137*, 2024.

[339] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.

[340] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[341] Bennett Landman et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Workshop*, 2015.

[342] Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. Pixeltone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2185–2194, 2013.

[343] Neal Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan, and Greg Ver Steeg. Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models. *arXiv preprint arXiv:2305.16597*, 2023.

[344] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*, 2021.

[345] Thanh-Dung Le, Ti Ti Nguyen, and Vu Nguyen Ha. The impact of lora adapters for llms on clinical nlp classification under data limitations. *arXiv preprint arXiv:2407.19299*, 2024.

[346] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

[347] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[348] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[349] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[350] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.

[351] Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Zhao, Yuexin Wu, Bo Li, et al. Conditional adapters: Parameter-efficient transfer learning with fast inference. *Advances in Neural Information Processing Systems*, 36:8152–8172, 2023.

[352] Xue-Liang Leng, Xiao-Ai Miao, and Tao Liu. Using recurrent neural network structure with enhanced multi-head self-attention for sentiment analysis. *Multimedia Tools and Applications*, 80:12581–12600, 2021.

[353] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[354] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[355] James Lester, Karl Branting, and Bradford Mott. Conversational agents. *The practical handbook of internet computing*, pages 220–240, 2004.

[356] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011.

[357] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[358] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021.

[359] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024.

[360] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

[361] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[362] Guochang Li, Chen Zhi, Jialiang Chen, Junxiao Han, and Shuiguang Deng. A comprehensive evaluation of parameter-efficient fine-tuning on automated program repair. *arXiv preprint arXiv:2406.05639*, 2024.

[363] Guochang Li, Chen Zhi, Jialiang Chen, Junxiao Han, and Shuiguang Deng. Exploring parameter-efficient fine-tuning of large language model on automated program repair. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 719–731, 2024.

[364] Haoling Li, Xin Zhang, Xiao Liu, Yeyun Gong, Yifan Wang, Yujiu Yang, Qi Chen, and Peng Cheng. Gradient-mask tuning elevates the upper limits of llm performance. *arXiv preprint arXiv:2406.15330*, 2024.

[365] Jonathan Li, Will Aitken, Rohan Bhambhoria, and Xiaodan Zhu. Prefix propagation: Parameter-efficient tuning for long sequences. *arXiv preprint arXiv:2305.12086*, 2023.

[366] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[367] Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–25, 2021.

[368] Juntao Li, Ling Luo, Tengxiao Lv, Chao Liu, Jiewei Qi, Zhihao Yang, Jian Wang, and Hongfei Lin. Instruction fine-tuning of large language models for traditional chinese medicine. In *China Conference on Knowledge Graph and Semantic Computing*, pages 419–430. Springer, 2024.

[369] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

[370] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdqunat: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007*, 2024.

[371] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[372] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[373] Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[374] Yi Li, Jinsong Wang, and Hongwei Zhang. A survey of state-of-the-art sharding blockchains: Models, components, and attack surfaces. *Journal of Network and Computer Applications*, 217:103686, 2023.

[375] Yuanchao Li, Yuan Gong, Chao-Han Huck Yang, Peter Bell, and Catherine Lai. Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction. *arXiv preprint arXiv:2409.15551*, 2024.

[376] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022.

[377] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

[378] Baohao Liao, Yan Meng, and Christof Monz. Parameter-efficient fine-tuning without introducing new latency. *arXiv preprint arXiv:2305.16742*, 2023.

[379] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809, 2018.

[380] Jeong-A Lim, Joohyun Lee, Jeongho Kwak, and Yeongjin Kim. Cutting-edge inference: Dynamic dnn model partitioning and resource scaling for mobile ai. *IEEE Transactions on Services Computing*, 2024.

[381] Bo Lin, Shangwen Wang, Zhongxin Liu, Yepang Liu, Xin Xia, and Xiaoguang Mao. Cct5: A code-change-oriented pre-trained model. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1509–1521, 2023.

[382] Diefan Lin, Yi Wen, Weishi Wang, and Yan Su. Enhanced sentiment intensity regression through lora fine-tuning on llama 3. *IEEE Access*, 2024.

[383] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023.

[384] Tzu-Han Lin, How-Shing Wang, Hao-Yung Weng, Kuang-Chen Peng, Zih-Ching Chen, and Hung-yi Lee. Peft for speech: Unveiling optimal placement, merging strategies, and ensemble techniques. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 705–709. IEEE, 2024.

[385] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

[386] Yang Lin, Xinyu Ma, Xu Chu, Yujie Jin, Zhibang Yang, Yasha Wang, and Hong Mei. Lora dropout as a sparsity regularizer for overfitting control. *arXiv preprint arXiv:2404.09610*, 2024.

[387] Yin Lin, Ziyang Wu, Qidong Huang, Xinran Liu, Baocai Yin, Jinshui Hu, Bing Yin, and Zengfu Wang. Efficient fine-tuning strategies for enhancing face recognition performance in challenging scenarios. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[388] Zheng Lin, Xuanjie Hu, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Ang Li, Praneeth Vepakomma, and Yue Gao. Splitlora: A split parameter-efficient fine-tuning framework for large language models. *arXiv preprint arXiv:2407.00952*, 2024.

[389] Zhisheng Lin, Han Fu, Chenghao Liu, Zhuo Li, and Jianling Sun. Pemt: Multi-task correlation guided mixture-of-experts enables parameter-efficient transfer learning. *arXiv preprint arXiv:2402.15082*, 2024.

[390] Bing Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[391] Can Liu, Kaijie Sun, Qingqing Zhou, Yuchen Duan, Jianhua Shu, Hongxing Kan, Zongyun Gu, and Jili Hu. Cpmi-chatglm: parameter-efficient fine-tuning chatglm with chinese patent medicine instructions. *Scientific reports*, 14(1):6403, 2024.

[392] Chun Liu, Hongguang Zhang, Kainan Zhao, Xinghai Ju, and Lin Yang. Llmembed: Rethinking lightweight llm's genuine function in text classification. *arXiv preprint arXiv:2406.03725*, 2024.

[393] Dongxu Liu, Bing Xu, Yinzhuo Chen, Bufan Xu, Wenpeng Lu, Muyun Yang, and Tiejun Zhao. Pmol: Parameter efficient moe for preference mixing of llm alignment. *arXiv preprint arXiv:2411.01245*, 2024.

[394] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. El evater: A benchmark for multimodal entity linking and visual attribute recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2443–2459. Association for Computational Linguistics, 2021.

[395] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[396] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[397] Jie Liu, Shaowei Chen, Bingquan Wang, Jiaxin Zhang, Na Li, and Tong Xu. Attention as relation: learning supervised multi-head self-attention for relation extraction. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 3787–3793, 2021.

[398] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 172–185. Springer, 2012.

[399] Linqing Liu and Xiaolong Xu. Self-attention mechanism at the token level: gradient analysis and algorithm optimization. *Knowledge-Based Systems*, 277:110784, 2023.

[400] Mingyuan Liu, Lu Xu, Shengnan Liu, and Jicong Zhang. Sparsity-and hybridity-inspired visual parameter-efficient fine-tuning for medical diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 627–637. Springer, 2024.

[401] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*, 2023.

[402] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114, 2024.

[403] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

[404] Tianci Liu, Zihan Dong, Linjun Zhang, Haoyu Wang, and Jing Gao. Mitigating heterogeneous token overfitting in llm knowledge editing. *arXiv preprint arXiv:2502.00602*, 2025.

[405] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023.

[406] Xiangyang Liu, Tianxiang Sun, Xuanjing Huang, and Xipeng Qiu. Late prompt tuning: A late prompt could be better than many prompts. *arXiv preprint arXiv:2210.11292*, 2022.

[407] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[408] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024.

[409] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14420–14430, 2023.

[410] Y Liu. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

[411] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[412] Yilun Liu, Yunpu Ma, Shuo Chen, Zifeng Ding, Bailan He, Zhen Han, and Volker Tresp. Perft: Parameter-efficient routed fine-tuning for mixture-of-expert model. *arXiv preprint arXiv:2411.08212*, 2024.

[413] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[414] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2646–2655, 2020.

[415] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[416] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[417] Carol Loganbill, Emily Hardy, and Ursula Delworth. Supervision: A conceptual model. *The counseling psychologist*, 10(1):3–42, 1982.

[418] Fei Long, Kai Zhou, and Weihua Ou. Sentiment analysis of text based on bidirectional lstm with multi-head attention. *Ieee Access*, 7:141960–141969, 2019.

[419] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008.

[420] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

[421] Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023.

[422] Junyi Lu, Lei Yu, Xiaojia Li, Li Yang, and Chun Zuo. Llama-reviewer: Advancing code review automation with large language models through parameter-efficient fine-tuning. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, pages 647–658. IEEE, 2023.

[423] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning, 2022.

[424] Tianxuan Lu, Jin Hu, and Pingping Chen. Benchmarking llama 3 for chinese news summation: Accuracy, cultural nuance, and societal value alignment. *Authorea Preprints*, 2024.

[425] Wenhao Lu, Jian Jiao, and Ruofei Zhang. Twinbert: Distilling knowledge to twin-structured bert models for efficient retrieval. corr abs/2002.06275 (2020). *arXiv preprint arXiv:2002.06275*, 2020.

[426] Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, et al. Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. *arXiv preprint arXiv:2305.15065*, 2023.

[427] Yuantong Lu and Zhanquan Wang. Few adjustable parameters prediction model based on lightweight prefix-tuning: Learning session dropout prediction model based on parameter-efficient prefix-tuning. *Applied Sciences*, 14(23):10772, 2024.

[428] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

[429] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan. Cdtb: A color and depth visual object tracking dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10013–10022, 2019.

[430] Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023.

[431] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[432] Xudong Luo, Zhiqi Deng, Binxia Yang, and Michael Y Luo. Pre-trained language models in medicine: A survey. *Artificial Intelligence in Medicine*, page 102904, 2024.

[433] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*, 2023.

[434] Fang Ma, Chen Zhang, Lei Ren, Jingang Wang, Qifan Wang, Wei Wu, Xiaojun Quan, and Dawei Song. Xprompt: Exploring the extreme of prompt tuning. *arXiv preprint arXiv:2210.04457*, 2022.

[435] Pingchuan Ma, Lennart Rietdorf, Dmytro Kotovenko, Vincent Tao Hu, and Björn Ommer. Does vlm classification benefit from llm description semantics? *arXiv preprint arXiv:2412.11917*, 2024.

[436] Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. Charbert: Character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*, 2020.

[437] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, Xiaopeng Hong, Yongjian Wu, and Rongrong Ji. Image captioning via dynamic path customization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[438] Yuhao Ma, Meina Kan, Shiguang Shan, and Xilin Chen. Learning deep face representation with long-tail data: An aggregate-and-disperse approach. *Pattern Recognition Letters*, 133:48–54, 2020.

[439] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[440] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[441] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on geoscience and remote sensing*, 55(2):645–657, 2016.

[442] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021.

[443] Joy Mahapatra and Utpal Garain. Impact of model size on fine-tuned llm performance in data-to-text generation: A state-of-the-art investigation. *arXiv preprint arXiv:2407.14088*, 2024.

[444] Huiyu Mai, Wenhao Jiang, and Zhihong Deng. Prefix-tuning based unsupervised text style transfer. *arXiv preprint arXiv:2310.14599*, 2023.

[445] Zhelu Mai, Jinran Zhang, Zhuoer Xu, and Zhaomin Xiao. Financial sentiment analysis meets llama 3: A comprehensive analysis. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, pages 171–175, 2024.

[446] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[447] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:2110.07577*, 2021.

[448] Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[449] Benjamin Markines, Ciro Cattuto, and Filippo Menczer. Social spam detection. In *Proceedings of the 5th international workshop on adversarial information retrieval on the web*, pages 41–48, 2009.

[450] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2728–2737. IEEE, 2021.

[451] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[452] Markus Maurer, J Christian Gerdes, Barbara Lenz, and Hermann Winner. *Autonomous driving: technical, legal and social aspects*. Springer Nature, 2016.

[453] Ian G McCarthy, Jaime Salcido, Joop Schaye, Juliana Kwan, Willem Elbers, Roi Kugel, Matthieu Schaller, John C Helly, Joey Braspenning, Carlos S Frenk, et al. The flamingo project: revisiting the s 8 tension and the role of baryonic physics. *Monthly Notices of the Royal Astronomical Society*, 526(4):5494–5519, 2023.

[454] Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 456–464, 2010.

[455] Kathleen McKeown. *Text generation*. Cambridge University Press, 1992.

[456] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[457] Julian Richard Medina and Jugal Kalita. Parallel attention mechanisms in neural machine translation. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 547–552. IEEE, 2018.

[458] Maryam Mehrabi, Abdelwahab Hamou-Lhadj, and Hossein Moosavi. The effectiveness of compact fine-tuned llms in log parsing. *Journal of Systems and Software*, 195(1):105320, 2024.

[459] Yelena Mejova. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*, 5:1–34, 2009.

[460] Xiangdi Meng, Damai Dai, Weiyao Luo, Zhe Yang, Shaoxiang Wu, Xiaochen Wang, Peiyi Wang, Qingxiu Dong, Liang Chen, and Zhifang Sui. Periodiclora: Breaking the low-rank bottleneck in lora optimization. *arXiv preprint arXiv:2402.16141*, 2024.

[461] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

[462] Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. Naamapadam: A large-scale named entity annotated data for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada, July 2023. Association for Computational Linguistics.

[463] Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Rui Kong, Xinyu Fang, Yazhu Dong, Yikang Zhang, Yunchun Li, Meng Li, et al. V-lora: An efficient and flexible system boosts vision applications with lora lmm. *arXiv preprint arXiv:2411.00915*, 2024.

[464] Xupeng Miao, Gabriele Oliaro, Xinhao Cheng, Mengdi Wu, Colin Unger, and Zhihao Jia. Flexllm: A system for co-serving large language model inference and parameter-efficient finetuning. *arXiv preprint arXiv:2402.18789*, 2024.

[465] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[466] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[467] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[468] Lyudmila Mihaylova, Tine Lefebvre, Herman Bruyninckx, Klaas Gadeyne, and Joris De Schutter. A comparison of decision making criteria and optimization methods for active robotic sensing. In *Numerical Methods and Applications: 5th International Conference, NMA 2002 Borovets, Bulgaria, August 20–24, 2002 Revised Papers 5*, pages 316–324. Springer, 2003.

[469] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[470] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[471] Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

[472] Arindam Mitra and Chitta Baral. Learning to use formulas to solve simple arithmetic problems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2144–2153, 2016.

[473] Rasoul Mojtahedzadeh, Abdelbaki Bouguerra, Erik Schaffernicht, and Achim J Lilienthal. Support relation analysis and decision making for safe robotic manipulation tasks. *Robotics and Autonomous Systems*, 71:99–117, 2015.

[474] Ali H Mokdad. The behavioral risk factors surveillance system: past, present, and future. *Annual review of public health*, 30(1):43–54, 2009.

[475] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898. IEEE, 2014.

[476] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

[477] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[478] Nusrat Nabi, Sumiya Islam, Sakib Ahmad Siddiquee, Syed Rakib Al Hossain, Mumenunnessa Keya, Sharun Akter Khushbu, and Md Sazzadur Ahamed. Sondhan: A comparative study of two proficiency language bangla-english on question-answer using attention mechanism. In *2021 31st International Conference on Computer Theory and Applications (ICCTA)*, pages 147–154. IEEE, 2021.

[479] Aarathi Rajagopalan Nair, Deepa Gupta, and B Premjith. Investigating translation for indic languages with bloomz-3b through prompting and lora fine-tuning. *Scientific Reports*, 14(1):24202, 2024.

[480] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, 2018.

[481] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. Pipedream: Generalized pipeline parallelism for dnn training. In *Proceedings of the 27th ACM symposium on operating systems principles*, pages 1–15, 2019.

[482] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–15, 2021.

[483] Stefan Escaida Navarro, Stephan Mühlbacher-Karrer, Hosam Alagi, Hubert Zangl, Keisuke Koyama, Björn Hein, Christian Duriez, and Joshua R Smith. Proximity perception in human-centered robotics: A survey on sensing systems and applications. *IEEE Transactions on Robotics*, 38(3):1599–1620, 2021.

[484] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. http://ufldl.stanford.edu/housenumbers/, 2011. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

[485] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. The street view house numbers (svhn) dataset. *Technical report*, 2018.

[486] Trong Thuan Nguyen, Hai Le, Truong Nguyen, Nguyen D Vo, and Khang Nguyen. A brief review of state-of-the-art object detectors on benchmark document images datasets. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(4):433–451, 2023.

[487] Jingchao Ni, Wei Cheng, Zhengzhang Chen, Takayoshi Asakura, Tomoya Soma, Sho Kato, and Haifeng Chen. Superclass-conditional gaussian mixture model for learning fine-grained embeddings. In *International Conference on Learning Representations*, 2021.

[488] Kun Nie, Tieju Ma, and Yoshiteru Nakamori. An approach to aid understanding emerging research fields—the case of knowledge management. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, 26(6):629–643, 2009.

[489] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

[490] Mahdi Nikdan, Soroush Tabesh, Elvir Crnčević, and Dan Alistarh. Rosa: Accurate parameter-efficient fine-tuning via robust adaptation. *arXiv preprint arXiv:2401.04679*, 2024.

[491] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2008.

[492] Tong Niu, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. Onealigner: Zero-shot cross-lingual transfer with one rich-resource language pair for low-resource sentence retrieval. *arXiv preprint arXiv:2205.08605*, 2022.

[493] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pages 71–80, 2020.

[494] Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*, 2020.

[495] Stuart F Oberman and Michael J Flynn. Implementing division and other floating-point operations: A system perspective. *MATHEMATICAL RESEARCH*, 90:18–24, 1996.

[496] Franz Josef Och. Giza++: Training of statistical translation models. *http://www. isi. edu/~ och/GIZA++. html*, 2001.

[497] Xianghan O'dea and Mike O'Dea. Is artificial intelligence really the next big thing in learning and teaching in higher education?: A conceptual paper. *Journal of University Teaching and Learning Practice*, 20(5):1–17, 2023.

[498] Peter F Oliva and George E Pawlas. *Supervision for today's schools*. ERIC, 2004.

[499] Aníbal Ollero and Luis Merino. Control and perception techniques for aerial robotics. *Annual reviews in Control*, 28(2):167–178, 2004.

[500] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[501] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9–16, 2019.

[502] Bei Ouyang, Shengyuan Ye, Liekang Zeng, Tianyi Qian, Jingyi Li, and Xu Chen. Pluto and charon: A time and memory efficient collaborative edge ai framework for personal llms fine-tuning. In *Proceedings of the 53rd International Conference on Parallel Processing*, pages 762–771, 2024.

[503] Lucas Page-Caccia, Edoardo Maria Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni. Multi-head adapter routing for cross-task generalization. *Advances in Neural Information Processing Systems*, 36:56916–56931, 2023.

[504] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[505] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023.

[506] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[507] Aditeya Pandey, Uzma Haque Syeda, Chaitya Shah, John A Guerra-Gomez, and Michelle A Borkin. A state-of-the-art survey of tasks for tree design and evaluation with a curated task dataset. *IEEE Transactions on Visualization and Computer Graphics*, 28(10):3563–3584, 2021.

[508] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia*, 17(11):2008–2020, 2015.

[509] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International journal of computer vision*, 38:15–33, 2000.

[510] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[511] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[512] Marinela Parović, Alan Ansell, Ivan Vulić, and Anna Korhonen. Cross-lingual transfer with target language-ready task adapters. *arXiv preprint arXiv:2306.02767*, 2023.

[513] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.

[514] Roberto Pecoraro, Valerio Basile, and Viviana Bono. Local multi-head channel self-attention for facial expression recognition. *Information*, 13(9):419, 2022.

[515] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*, 2022.

[516] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[517] Miao Peng, Ben Liu, Wenjie Xu, Zihao Jiang, Jiahui Zhu, and Min Peng. Deja vu: Contrastive historical modeling with prefix-tuning for temporal knowledge graph reasoning. *arXiv preprint arXiv:2404.00051*, 2024.

[518] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019. 126-class subset of DomainNet with 586,575 images across 6 domains.

[519] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[520] Zhiyuan Peng, Xuyang Wu, Qifan Wang, Sravanthi Rajanala, and Yi Fang. Q-peft: Query-dependent parameter efficient fine-tuning for text reranking with large language models. *arXiv preprint arXiv:2404.04522*, 2024.

[521] John Peterson. Policy networks. Technical report, Institut für Höhere Studien, 2003.

[522] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

[523] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020.

[524] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020.

[525] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020-not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1525–1535, 2022.

[526] Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[527] Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. Focused attention improves document-grounded generation. *arXiv preprint arXiv:2104.12714*, 2021.

[528] Nusrat Jahan Prottasha, Md Kowsher, Hafijur Raman, Israt Jahan Anny, Prakash Bhat, Ivan Garibay, and Ozlem Garibay. User profile with large language models: Construction, updating, and benchmarking. *arXiv preprint arXiv:2502.10660*, 2025.

[529] Nusrat Jahan Prottasha, Asif Mahmud, Md Shohanur Islam Sobuj, Prakash Bhat, Md Kowsher, Niloofar Yousefi, and Ozlem Ozmen Garibay. Parameter-efficient fine-tuning of large language models using semantic knowledge tuning. *Scientific Reports*, 14(1):30667, 2024.

[530] Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud, and Mohammed Baz. Transfer learning for sentiment analysis using bert based supervised fine-tuning. *Sensors*, 22(11):4157, 2022.

[531] George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. Empirical analysis of the strengths and weaknesses of peft techniques for llms. *arXiv preprint arXiv:2304.14999*, 2023.

[532] Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Jing Yi, Weize Chen, Zhiyuan Liu, Juanzi Li, Lei Hou, et al. Exploring universal intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*, 2021.

[533] Bo Qu, Gao Cong, Cuiping Li, Aixin Sun, and Hong Chen. An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, 63(5):889–903, 2012.

[534] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[535] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[536] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

[537] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[538] Hamed Rahimi, Mouad Abrini, Mahdi Khoramshahi, and Mohamed Chetouani. User-vlm: Llm contextualization with multimodal pre-trained user models. *Under Review*, 2025.

[539] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

[540] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[541] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[542] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1092–1096, 2016.

[543] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[544] Bradley C Reed, Jesslyn F Brown, Darrel VanderZee, Thomas R Loveland, James W Merchant, and Donald O Ohlen. Measuring phenological variability from satellite imagery. *Journal of vegetation science*, 5(5):703–714, 1994.

[545] Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. M2D2: A massively multi-domain language modeling dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 964–975, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[546] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[547] Stefan Emil Repede and Remus Brad. Llama 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers (2073-431X)*, 13(11), 2024.

[548] Moritz Reuss and Rudolf Lioutikov. Multimodal diffusion transformer for learning from play. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[549] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*, 2024.

[550] RAW Rhodes. Policy networks. *The Oxford handbook of public policy*, 6:425–447, 2006.

[551] Arnaud Rosay, Florent Carlier, and Pascal Leroux. Feed-forward neural network for network intrusion detection. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–6. IEEE, 2020.

[552] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, 2004.

[553] Mikhail M Rovnyagin, Dmitry M Sinelnikov, Artem A Eroshev, Tatyana A Rovnyagina, and Alexander V Tikhomirov. Optimizing cache memory usage methods for chat llm-models in paas installations. In *2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon)*, pages 277–280. IEEE, 2024.

[554] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[555] Victoria L Rubin, Yimin Chen, and Lynne Marie Thorimbert. Artificially intelligent conversational agents in libraries. *Library Hi Tech*, 28(4):496–522, 2010.

[556] Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*, 2020.

[557] Bharat Runwal, Tejaswini Pedapati, and Pin-Yu Chen. From peft to deft: Parameter efficient finetuning for reducing activation density in transformers. *arXiv preprint arXiv:2402.01911*, 2024.

[558] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[559] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[560] Mingi Ryu. [RE] ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Unpublished Manuscript*, 2021. https://github.com/dkssud8150/RE-ALBERT.

[561] Dov Sagi and Bela Julesz. " where" and" what" in vision. *Science*, 228(4704):1217–1219, 1985.

[562] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[563] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[564] Alireza Salemi and Hamed Zamani. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *arXiv preprint arXiv:2409.09510*, 2024.

[565] David Sanders. Perception in robotics. *Industrial Robot: An International Journal*, 26(2), 1999.

[566] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050, 2003.

[567] V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[568] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[569] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[570] Iqbal H Sarker. Ai-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN computer science*, 3(2):158, 2022.

[571] Md Shahriare Satu, Md Hasnat Parvez, et al. Review of integrated applications with aiml based chatbot. In *2015 International Conference on Computer and Information Engineering (ICCIE)*, pages 87–90. IEEE, 2015.

[572] Eric Saund, David Fleet, Daniel Larner, and James Mahoney. Perceptually-supported image editing of text and graphics. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 183–192, 2003.

[573] Saumya Saxena, Mohit Sharma, and Oliver Kroemer. Multi-resolution sensing for real-time control with vision-language models. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[574] Dana Schaa and David Kaeli. Exploring the multiple-gpu design space. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pages 1–12. IEEE, 2009.

[575] Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R Zaiane, and Matthew E Taylor. The evolving landscape of llm-and vlm-integrated reinforcement learning. *arXiv preprint arXiv:2502.15214*, 2025.

[576] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[577] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[578] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[579] Arijit Sehanobish, Avinava Dubey, Krzysztof Choromanski, Somnath Basu Roy Chowdhury, Deepali Jain, Vikas Sindhwani, and Snigdha Chaturvedi. Structured unrestricted-rank matrices for parameter efficient fine-tuning. *arXiv preprint arXiv:2406.17740*, 2024.

[580] Md Shafikuzzaman, Md Rakibul Islam, Alex C Rolli, Sharmin Akhter, and Naeem Seliya. An empirical evaluation of the zero-shot, few-shot, and traditional fine-tuning based pretrained language models for sentiment analysis in software engineering. *IEEE Access*, 2024.

[581] Shahriar Shakil, Atik Asif Khan Akash, Nusrat Nabi, Mahmudul Hasan, and Aminul Haque. Pithanet: a transfer learning-based approach for traditional pitha classification. *International Journal of Electrical & Computer Engineering (2088-8708)*, 13(5), 2023.

[582] Madhu Sharma and Cmaune Sharma. A review on diverse applications of case-based reasoning. *Advances in computing and intelligent systems: Proceedings of ICACM 2019*, pages 511–517, 2020.

[583] Mohit Sharma, Claudio Fantacci, Yuxiang Zhou, Skanda Koppula, Nicolas Heess, Jon Scholz, and Yusuf Aytar. Lossless adaptation of pretrained vision models for robotic manipulation. *arXiv preprint arXiv:2304.06600*, 2023.

[584] Junhong Shen, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolo Fusi. Tag-llm: Repurposing general-purpose llms for specialized domains. *arXiv preprint arXiv:2402.05140*, 2024.

[585] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[586] Zhengpeng Shi and Haoran Luo. Cre-llm: A domain-specific chinese relation extraction framework with fine-tuned large language model. *arXiv preprint arXiv:2404.18085*, 2024.

[587] Zhengxiang Shi and Aldo Lipani. Dept: Decomposed prompt tuning for parameter-efficient fine-tuning. *arXiv preprint arXiv:2309.05173*, 2023.

[588] K Kirk Shung, Michael Smith, and Benjamin MW Tsui. *Principles of medical imaging*. Academic Press, 2012.

[589] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

[590] Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. A three-pronged approach to cross-lingual adaptation with multilingual llms. *arXiv preprint arXiv:2406.17377*, 2024.

[591] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021.

[592] Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Rahul Dodhia, Juan Lavista Ferres, and Bonnie Berger. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 121(26):e2405840121, 2024.

[593] Tony C Smith et al. Semantic role labeling via instance-based learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 180–188, 2006.

[594] Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25:5–35, 2005.

[595] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[596] Jesus Solano, Mardhiyah Sanni, Oana-Maria Camburu, and Pasquale Minervini. Sparsefit: Few-shot prompting with sparse fine-tuning for jointly generating predictions and natural language explanations. *arXiv preprint arXiv:2305.13235*, 2023.

[597] Chao Song, Zhihao Ye, Qiqiang Lin, Qiuying Peng, and Jun Wang. A framework to implement 1+ n multi-task fine-tuning pattern in llms using the cgc-lora algorithm. *arXiv preprint arXiv:2402.01684*, 2024.

[598] Haobo Song, Hao Zhao, Soumajit Majumder, and Tao Lin. Increasing model capacity for free: A simple strategy for parameter efficient fine-tuning. *arXiv preprint arXiv:2407.01320*, 2024.

[599] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[600] Wongkot Sriurai. Improving text categorization by using a topic model. *Advanced Computing*, 2(6):21, 2011.

[601] M. Steyvers, H. Tejeda, A. Kumar, et al. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231, 2025.

[602] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[603] Junda Su, Zirui Liu, Zeju Qiu, Weiyang Liu, and Zhaozhuo Xu. In defense of structural sparse adapters for concurrent llm serving. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*, 2024.

[604] Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. On transferability of prompt tuning for natural language processing. *arXiv preprint arXiv:2111.06719*, 2021.

[605] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[606] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

[607] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

[608] Jothi Prasanna Shanmuga Sundaram, Wan Du, and Zhiwei Zhao. A survey on lora networking: Research problems, current solutions, and open issues. *IEEE Communications Surveys & Tutorials*, 22(1):371–388, 2019.

[609] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022.

[610] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021.

[611] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. https://arxiv.org/abs/2210.09261, 2022. arXiv:2210.09261.

[612] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

[613] Anushka Swarup, Avanti Bhandarkar, Olivia P Dizon-Paradis, Ronald Wilson, and Damon L Woodard. Maximizing relation extraction potential: A data-centric study to unveil challenges and opportunities. *IEEE Access*, 2024.

[614] Zar Nawab Khan Swati, Qinghua Zhao, Muhammad Kabir, Farman Ali, Zakir Ali, Saeed Ahmed, and Jianfeng Lu. Brain tumor classification for mr images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75:34–46, 2019.

[615] Maite Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347, 2016.

[616] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[617] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[618] Zeqi Tan, Yongliang Shen, Xiaoxia Cheng, Chang Zong, Wenqi Zhang, Jian Shao, Weiming Lu, and Yueting Zhuang. Learning global controller in latent space for parameter-efficient fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4044–4055, 2024.

[619] Ningyuan Tang, Minghao Fu, Ke Zhu, and Jianxin Wu. Low-rank attention side-tuning for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04009*, 2024.

[620] Yiwen Tang, Ray Zhang, Zoey Guo, Xianzheng Ma, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5171–5179, 2024.

[621] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[622] Oguzhan Tas and Farzad Kiyani. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213, 2007.

[623] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10, 2009.

[624] Niall Taylor, Upamanyu Ghose, Omid Rohanian, Mohammadmahdi Nouriborji, Andrey Kormilitzin, David Clifton, and Alejo Nevado-Holgado. Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks. *arXiv preprint arXiv:2402.10597*, 2024.

[625] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[626] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[627] GLM Team, A Zeng, B Xu, et al. Chatglm: a family of large language models from glm-130b to glm-4 all tools. arxiv e-prints. *arXiv preprint arXiv:2406.12793*, 2024.

[628] Anshul Thakur, Vinayak Abrol, Pulkit Sharma, Tingting Zhu, and David A Clifton. Incremental trainable parameter selection in deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[629] William Thies, Vikram Chandrasekhar, and Saman Amarasinghe. A practical approach to exploiting coarse-grained pipeline parallelism in c programs. In *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*, pages 356–369. IEEE, 2007.

[630] Allwyn Bat Thomas, Ananya Reetha Noble, Anna Wilson, Leya Elizabeth Sunny, and Rini Thazhathoot Paul. Exploring lora for parameter-efficient fine-tuning of llms in enhanced algorithm-to-python-source-code translation task. In *AIP Conference Proceedings*, volume 3280. AIP Publishing, 2025.

[631] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer, 2020.

[632] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.

[633] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[634] Hanh Thi Hong Tran, Nishan Chatterjee, Senja Pollak, and Antoine Doucet. Deberta beats behemoths: A comparative analysis of fine-tuning, prompting, and peft approaches on legallensner. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 371–380, 2024.

[635] V Javier Traver and Alexandre Bernardino. A review of log-polar imaging for visual perception in robotics. *Robotics and Autonomous Systems*, 58(4):378–398, 2010.

[636] H Trung. Multimodal dialogue management-state of the art. *Human Media Interaction Department, University of Twente*, 2:32, 2006.

[637] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[638] Oleksandr Tsymbal, Artem Bronnikov, and Andriy Yerokhin. Adaptive decision-making for robotic tasks. In *2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL)*, pages 594–597. IEEE, 2019.

[639] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7725–7735, 2023.

[640] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 329–341, 2020.

[641] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *arXiv preprint arXiv:1908.04616*, 2019.

[642] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[643] Maria Valera and Sergio A Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings-Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[644] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*, 2022.

[645] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604. IEEE, 2015.

[646] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[647] Neelay Velingker, Jason Liu, Amish Sethi, William Dodds, Zhiqiu Xu, Saikat Dutta, Mayur Naik, and Eric Wong. Clam: Unifying finetuning, quantization, and pruning by chaining llm adapter modules. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*, 2024.

[648] Vinoth Kumar Venkatesan, Mahesh Thyluru Ramakrishna, Anatoliy Batyuk, Andrii Barna, and Bohdana Havrysh. High-performance artificial intelligence recommendation of quality research papers using effective collaborative approach. *Systems*, 11(2):81, 2023.

[649] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

[650] L VISION and SACRI TODAY'S. Demystifying the development of an organizational vision. *Sloan management review*, 1996.

[651] Kushala VM, Harikrishna Warrier, Yogesh Gupta, et al. Fine tuning llm for enterprise: Practical guidelines and recommendations. *arXiv preprint arXiv:2404.10779*, 2024.

[652] Ellen M Voorhees et al. Overview of the trec 2003 robust retrieval track. In *Trec*, pages 69–77, 2003.

[653] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*, 2021.

[654] Danilo Vucetic, Mohammadreza Tayaranian, Maryam Ziaeefard, James J Clark, Brett H Meyer, and Warren J Gross. Efficient fine-tuning of bert models on the edge. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1838–1842. IEEE, 2022.

[655] Ayham Wael and Amer Madi. Accelerating artificial intelligence: The role of gpus in deep learning and computational advancements. *East Journal of Engineering*, 1(1):31–46, 2025.

[656] Marcel Wagenländer, Guo Li, Bo Zhao, Luo Mai, and Peter Pietzuch. Tenplex: Dynamic parallelism for deep learning using parallelizable tensor collections. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, pages 195–210, 2024.

[657] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[658] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. Ace 2005 multilingual training corpus. *(No Title)*, 2006.

[659] Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840, 2004.

[660] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[661] Alex Wang, Ian F Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R Thomas McCoy, Roma Patel, et al. jiant 1.2: A software toolkit for research on general-purpose text understanding models. *Note: http://jiant. info/Cited by: footnote*, 4, 2019.

[662] Alex Hai Wang. Don't follow me: Spam detection in twitter. In *2010 international conference on security and cryptography (SECRYPT)*, pages 1–10. IEEE, 2010.

[663] Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[664] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017.

[665] Boxiang Wang, Qifan Xu, Zhengda Bian, and Yang You. Tesseract: Parallelize the tensor parallelism efficiently. In *Proceedings of the 51st International Conference on Parallel Processing*, pages 1–11, 2022.

[666] De Wang, Danesh Irani, and Calton Pu. A social-spam detection framework. In *Proceedings of the 8th annual collaboration, electronic messaging, anti-abuse and Spam conference*, pages 46–54, 2011.

[667] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

[668] Guanhua Wang, Olatunji Ruwase, Bing Xie, and Yuxiong He. Fastpersist: Accelerating model checkpointing in deep learning. *arXiv preprint arXiv:2406.13768*, 2024.

[669] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. Progress in machine translation. *Engineering*, 18:143–153, 2022.

[670] Haixin Wang, Jianlong Chang, Yihang Zhai, Xiao Luo, Jinan Sun, Zhouchen Lin, and Qi Tian. Lion: Implicit vision prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5372–5380, 2024.

[671] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[672] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

[673] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[674] Haowen Wang, Tao Sun, Kaixiang Ji, Jian Wang, Cong Fan, and Jinjie Gu. Orchmoe: Efficient multi-adapter learning with task-skill synergy. *arXiv preprint arXiv:2401.10559*, 2024.

[675] Haoyu Wang, Tianci Liu, Ruirui Li, Monica Cheng, Tuo Zhao, and Jing Gao. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. *arXiv preprint arXiv:2406.10777*, 2024.

[676] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[677] Heyuan Wang, Ziyi Wu, and Junyu Chen. Multi-turn response selection in retrieval-based chatbots with iterated attentive convolution matching network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1081–1090, 2019.

[678] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1788–1797, 2018.

[679] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7512–7520, 2018.

[680] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[681] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017.

[682] Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5517–5525, 2024.

[683] Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, et al. Aprompt: Attention prompt tuning for efficient adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9147–9160, 2023.

[684] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[685] Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Trans-lora: Towards data-free transferable parameter efficient finetuning. *arXiv preprint arXiv:2405.17258*, 2024.

[686] Shuai Wang and Zhendong Su. Metamorphic testing for object detection systems. *arXiv preprint arXiv:1912.12162*, 2019.

[687] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.

[688] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 2025.

[689] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2023.

[690] Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*, 2023.

[691] Yang Wang and Chenghua Lin. Tougher text, smarter models: Raising the bar for adversarial defence benchmarks. *arXiv preprint arXiv:2501.02654*, 2025.

[692] Yaqin Wang, Jin Wei-Kocsis, John A Springer, and Eric T Matson. Deep learning in audio classification. In *International Conference on Information and Software Technologies*, pages 64–77. Springer, 2022.

[693] Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*, 2022.

[694] Yu Emma Wang, Gu-Yeon Wei, and David Brooks. Benchmarking tpu, gpu, and cpu platforms for deep learning. *arXiv preprint arXiv:1907.10701*, 2019.

[695] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*, 2021.

[696] Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861*, 2023.

[697] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2613–2623, 2022.

[698] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in neural information processing systems*, 35:14388–14402, 2022.

[699] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[700] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 6:625–641, 2018.

[701] Pengfei Wei, Yiping Ke, and Chi Keong Goh. A general domain specific feature transfer framework for hybrid domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1440–1451, 2018.

[702] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2):449–460, 2016.

[703] Ralph Weischedel et al. Ontonotes release 5.0. LDC2013T19, 2013. 2.9M words, multilingual corpus with NER and coreference annotations.

[704] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046, 2022.

[705] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210, 2005.

[706] Kimball Wiles and John T Lovell. *Supervision for Better Schools*. ERIC, 1975.

[707] Victor Wiley and Thomas Lucas. Computer vision and image processing: a paper review. *International journal of artificial intelligence research*, 2(1):29–36, 2018.

[708] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[709] Alphus D Wilson. Diverse applications of electronic-nose technologies in agriculture and forestry. *Sensors*, 13(2):2295–2348, 2013.

[710] Andrew P Witkin and Jay M Tenenbaum. On the role of structure in vision. In *Human and machine vision*, pages 481–543. Elsevier, 1983.

[711] Dieter Wolff. Integrating language and content in the language classroom: Are transfer of knowledge and of language ensured? *ASp. la revue du GERAS*, 41-42:35–46, 2003.

[712] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[713] Workspace. Oxford flowers 102 dataset. https://universe.roboflow.com/workspace-iiqtq/oxford-flowers-102, jul 2023. Visited on 2025-03-28.

[714] Maurice Wright. Policy community, policy network and comparative industrial policies. *Political Studies*, 36(4):593–612, 1988.

[715] Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36:61060–61084, 2023.

[716] Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[717] Muling Wu, Wenhao Liu, Jianhan Xu, Changze Lv, Zixuan Ling, Tianlong Li, Longtao Huang, Xiaoqing Zheng, and Xuan-Jing Huang. Parameter efficient multi-task fine-tuning by learning to transfer token-wise prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8734–8746, 2023.

[718] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[719] Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364, 2019.

[720] Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. *arXiv preprint arXiv:2406.11909*, 2024.

[721] Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974, 2023.

[722] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[723] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[724] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, VG Vydiswaran, and Hao Ma. Idpg: An instance-dependent prompt generation method. *arXiv preprint arXiv:2204.04497*, 2022.

[725] Markus Wulfmeier, Arunkumar Byravan, Tim Hertweck, Irina Higgins, Ankush Gupta, Tejas Kulkarni, Malcolm Reynolds, Denis Teplyashin, Roland Hafner, Thomas Lampe, et al. Representation matters: Improving perception and exploration for robotics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6512–6519. IEEE, 2021.

[726] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010.

[727] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[728] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[729] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023.

[730] Ming Xie. *Fundamentals of robotics: linking perception to action*, volume 54. World Scientific Publishing Company, 2003.

[731] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831, 2012.

[732] Tianfang Xie, Tianjing Li, Wei Zhu, Wei Han, and Yi Zhao. Pedro: Parameter-efficient fine-tuning with prompt dependent representation modification. *arXiv preprint arXiv:2409.17834*, 2024.

[733] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16085–16093, 2024.

[734] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024.

[735] Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885, 2024.

[736] Chengming Xu, Siqian Yang, Yabiao Wang, Zhanxiong Wang, Yanwei Fu, and Xiangyang Xue. Exploring efficient few-shot adaptation for vision transformers. *arXiv preprint arXiv:2301.02419*, 2023.

[737] Haoran Xu, Benjamin Van Durme, and Kenton Murray. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. *arXiv preprint arXiv:2109.04588*, 2021.

[738] Haoyun Xu, Runzhe Zhan, Derek F Wong, and Lidia S Chao. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. *arXiv preprint arXiv:2403.11621*, 2024.

[739] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023.

[740] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023.

[741] Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. Fwdllm: Efficient fedllm using forward gradient. *arXiv preprint arXiv:2308.13894*, 2023.

[742] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*, 2021.

[743] Zhaozhuo Xu, Zirui Liu, Beidi Chen, Shaochen Zhong, Yuxin Tang, WANG Jue, Kaixiong Zhou, Xia Hu, and Anshumali Shrivastava. Soft prompt recovers compressed llms, transferably. In *Forty-first International Conference on Machine Learning*, 2021.

[744] Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. An empirical study on cross-lingual vocabulary adaptation for efficient generative llm inference. *arXiv preprint arXiv:2402.10712*, 2024.

[745] Haibin Yan, Marcelo H Ang, and Aun Neow Poo. A survey on perception methods for human–robot interaction in social robots. *International Journal of Social Robotics*, 6:85–119, 2014.

[746] Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

[747] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[748] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173. IEEE, 2013.

[749] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641, 2003.

[750] Juncheng Yang, Zuchao Li, Shuai Xie, Weiping Zhu, Wei Yu, and Shijun Li. Cross-modal adapter: Parameter-efficient transfer learning approach for vision-language models. *arXiv preprint arXiv:2404.12588*, 2024.

[751] Xiaocong Yang, James Y Huang, Wenxuan Zhou, and Muhao Chen. Parameter-efficient tuning with special token adaptation. *arXiv preprint arXiv:2210.04382*, 2022.

[752] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[753] Kai Yao, Penglei Gao, Lichun Li, Yuan Zhao, Xiaofeng Wang, Wei Wang, and Jianke Zhu. Layer-wise importance matters: Less memory for better performance in parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2410.11772*, 2024.

[754] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017.

[755] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[756] He Ye, Matias Martinez, and Martin Monperrus. Neural program repair with execution-based backpropagation. in 2022 ieee/acm 44th international conference on software engineering (icse). *IEEE, 1506ś1518*, 2022.

[757] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

[758] Yelp Inc. Yelp dataset challenge round 7. https://www.yelp.com/dataset, 2015. 700,000 business reviews with sentiment labels.

[759] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[760] Esma Yildirim, Engin Arslan, Jangyoung Kim, and Tevfik Kosar. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. *IEEE Transactions on Cloud Computing*, 4(1):63–75, 2015.

[761] Dongshuo Yin, Xueting Han, Bin Li, Hao Feng, and Jing Bai. Parameter-efficient is not sufficient: Exploring parameter, memory, and time efficient adapter tuning for dense predictions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1398–1406, 2024.

[762] Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *arXiv preprint arXiv:2406.01563*, 2024.

[763] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[764] Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

[765] Fu-Hao Yu, Kuan-Yu Chen, and Ke-Han Lu. Non-autoregressive asr modeling using pre-trained language models for chinese speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1474–1482, 2022.

[766] Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an llm? a new benchmark dataset and approach for detecting ai text in peer review. *arXiv preprint arXiv:2502.19614*, 2025.

[767] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[768] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[769] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[770] Zhengqing Yuan, Huiwen Xue, Xinyi Wang, Yongming Liu, Zhuanzhe Zhao, and Kun Wang. Artgpt-4: artistic vision-language understanding with adapter-enhanced minigpt-4. *arXiv preprint arXiv:2305.07490*, 19, 2023.

[771] Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. Building a family of data augmentation models for low-cost llm fine-tuning on the cloud. *arXiv preprint arXiv:2412.04871*, 2024.

[772] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[773] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

[774] Mohamad Hasan Zahweh, Hasan Nasrallah, Mustafa Shukor, Ghaleb Faour, and Ali J Ghandour. Empirical study of peft techniques for winter-wheat segmentation. *Environmental Sciences Proceedings*, 29(1):50, 2023.

[775] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

[776] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722. IEEE, 2018.

[777] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024.

[778] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[779] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[780] Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14161–14170, 2023.

[781] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[782] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

[783] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[784] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[785] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 698–714. Springer, 2020.

[786] Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. Personalized llm response generation with parameterized memory injection. *arXiv preprint arXiv:2404.03565*, 2024.

[787] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[788] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

[789] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[790] Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. Autolora: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. *arXiv preprint arXiv:2403.09113*, 2024.

[791] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[792] Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. Beyond the turn-based game: Enabling real-time conversations with duplex models. *arXiv preprint arXiv:2406.15718*, 2024.

[793] Y Zhang. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[794] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[795] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[796] Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. Towards adaptive prefix tuning for parameter-efficient language model fine-tuning. *arXiv preprint arXiv:2305.15212*, 2023.

[797] Zhengkun Zhang, Wenya Guo, Xiaojun Meng, Yasheng Wang, Yadao Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. Hyperpelt: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. *arXiv preprint arXiv:2203.03878*, 2022.

[798] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.

[799] Bowen Zhao, Hannaneh Hajishirzi, and Qingqing Cao. Apt: Adaptive pruning and tuning pre-trained language models for efficient training and inference. *arXiv preprint arXiv:2401.12200*, 2024.

[800] Hao Zhao, Jie Fu, and Zhaofeng He. Prototype-based hyperadapter for sample-efficient multi-task tuning. *arXiv preprint arXiv:2310.11670*, 2023.

[801] Hongyu Zhao, Hao Tan, and Hongyuan Mei. Tiny-attention adapter: Contexts are more important than the number of parameters. *arXiv preprint arXiv:2211.01979*, 2022.

[802] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[803] Wentao Zhao, Jiaming Chen, Ziyu Meng, Donghui Mao, Ran Song, and Wei Zhang. Vlmpc: Vision-language model predictive control for robotic manipulation. *arXiv preprint arXiv:2407.09829*, 2024.

[804] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[805] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10394–10403, 2019.

[806] Léon Zheng. *Data frugality and computational efficiency in deep learning*. PhD thesis, Ecole normale supérieure de lyon-ENS LYON, 2024.

[807] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[808] Xianrui Zheng, Chao Zhang, and Philip C Woodland. Adapting gpt, gpt-2 and bert language models for speech recognition. In *2021 IEEE Automatic speech recognition and understanding workshop (ASRU)*, pages 162–168. IEEE, 2021.

[809] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017.

[810] Aihua Zhou, Yujun Ma, Wanting Ji, Ming Zong, Pei Yang, Min Wu, and Mingzhe Liu. Multi-head attention-based two-stream efficientnet for action recognition. *Multimedia Systems*, 29(2):487–498, 2023.

[811] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[812] GuangXu Zhou, Hemant Joshi, and Coskun Bayrak. Topic categorization for relevancy and opinion detection. In *TREC*, 2007.

[813] Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. Autopeft: Automatic configuration search for parameter-efficient fine-tuning. *Transactions of the Association for Computational Linguistics*, 12:525–542, 2024.

[814] Nan Zhou, Huiqun Wang, Yaoyan Zheng, and Di Huang. Progressive parameter efficient transfer learning for semantic segmentation. In *The Thirteenth International Conference on Learning Representations*, 2024.

[815] Shibin Zhou, Kan Li, and Yushu Liu. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409, 2009.

[816] Xiongtao Zhou, Jie He, Yuhua Ke, Guangyao Zhu, Víctor Gutiérrez-Basulto, and Jeff Z Pan. An empirical study on parameter-efficient fine-tuning for multimodal large language models. *arXiv preprint arXiv:2406.05130*, 2024.

[817] Yancong Zhou, Xuemei Zhang, Yan Wang, and Bo Zhang. Transfer learning and its application research. In *Journal of Physics: Conference Series*, volume 1920, page 012058. IOP Publishing, 2021.

[818] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[819] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[820] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9516–9526, 2023.

[821] Qihao Zhu, Zeyu Sun, Yuan-an Xiao, Wenjie Zhang, Kang Yuan, Yingfei Xiong, and Lu Zhang. A syntax-guided edit decoder for neural program repair. In *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 341–353, 2021.

[822] Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. Counter-interference adapter for multilingual machine translation. *arXiv preprint arXiv:2104.08154*, 2021.

[823] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[824] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

Dataset	Domain	Year	Task Type	I/O Format	Size	Train/Test
20 Newsgroups [471]	NLP	1999	Text Classification	Text → Category	20K	11K / 7.5K
CoNLL03 [566]	NLP	2003	NER	Text → Entity tags	2,302	946 / 231
CoNLL04 [552]	NLP	2004	Relation extraction	Text → BIO tags	1,437	8,936 / 1,671
Caltech101 [167]	CV	2004	Image Classification	Images → Class	9,144	-
SST-2 [595]	NLP	2005	Sentiment Analysis	Text → Binary	215K	67K / 1.8K
MRPC [140]	NLP	2005	Paraphrase Detection	Sentence pairs → Binary	5,800	3,669 / 1,724
ACE2005 [658]	NLP	2005	Information Extraction	Text → Entities	599	529 / 40
SRL WSJ & Brown [65]	NLP	2005	Semantic Role Labeling	Text → Predicate-args	75K	-
PASCAL VOC [159]	CV	2005–12	Object Detection	Images → Boxes/labels	11,530	5,717
WMT [319]	NLP	2006-	Machine Translation	Text → Text	-	-
MPQA [705]	NLP	2006	Opinion Mining	Text → Text	535	-
Common Crawl [172]	NLP	2008	NLP Pretraining	Web text → Text	386 TiB	-
Oxford Flowers [491]	CV	2008	Fine-grained Class.	Images → Labels	8,189	7.17K / 1.02K
Common Crawl [172]	NLP	2008	NLP Pretraining	Web text → Text	386 TiB	-
CIFAR-10/100 [331]	CV	2009	Image Classification	RGB → Labels	60K	50K / 10K
STL-10 [110]	CV	2011	Image Classification	RGB → Labels	13K	5K / 8K
SUN397 [726]	CV	2010	Scene Classification	RGB → Labels	108,753	-
i2b2 2010 RE [642]	NLP	2010	Relation Extraction	Clinical → Relations	877	394 / 477
Semeval-2010 [314]	NLP	2010	Semantic Role	Text → Labels	284	144 / 100
SBU [500]	CV	2011	Image Captioning	Images → Captions	1M	-
COPA [190]	NLP	2011	Causal reasoning	Premise → Answer	1,000	500 / 500
IMDb [440]	NLP	2011	Sentiment analysis	Text → Binary	50K	25K / 25K
CUB-200-2011 [657]	CV	2011	Fine-grained Class.	Images → Labels	11,788	5,994 / 5,794
Stanford Dogs [398]	CV	2011	Fine-Grained Class.	Images → Labels	20,580	14.4K / 6.17K
SVHN [485]	CV	2011	Digit recognition	Images → Digits	600K	-
OxfordPets [511]	CV	2012	Fine-Grained Class.	RGB → 37 breeds	7,390	-
ImageNet1K [131]	CV	2012-17	Image classification	RGB → 1000 classes	1.43M	1.28M / 100K
WebQ [38]	NLP	2013	Question Answering	Questions → KB Entity	6,642	3,778 / 2,032
OntoNotes [703]	NLP	2013	NER, Coreference	Text → BIO-format	2.95M	-
DUT [748]	CV	2013	Salient Detection	RGB → Binary masks	5,168	-
Yelp Polarity [758]	NLP	2015	Sentiment Analysis	Text Reviews → Sentiment Labels	560K	38K
CNN/DailyMail [578]	NLP	2015	Abstractive Summarization	News Articles → Summaries	313K	287K
WebNLG [708]	NLP	2017	NLG	RDF → Text	13.2K	-
STS-B [66]	NLP	2017	Similarity	Sent. pairs → Score	8.6K	5.7K / 1.4K
HS [128]	NLP	2017	Hate Speech Detection	Tweets → Labels	25K	19.8K / 4.8K
WikiSQL [809]	NLP	2017	Text-to-SQL	Query + Table → SQL	80.7K	56.4K / 15.9K
ARC [106]	NLP	2018	MC QA	Q + MC → Answer	7.8K	-
GLUE [660]	NLP	2018	NLU	Sentences → Labels	-	-
OBQA [467]	NLP	2018	Commonsense QA	MC → Answer	6K	5K / 500
ARC-Easy [106]	NLP	2018	MC QA	MC → Answer	5.2K	2.3K / 2.4K
ARC-Challenge [106]	NLP	2018	MC QA	MC → Answer	2.6K	1.1K / 1.2K
MultiRC [304]	NLP	2018	Multi-Sent. RC	Para. → MC	10K	6K
XSum [480]	NLP	2018	Abstr. Summar.	News → Summ.	226.7K	204K / 11.3K
CoLA [700]	NLP	2018	Sent. Accept.	Sent. → Binary	10.7K	8.6K
UCF101 [599]	Vision	2018	Action Rec.	Video → Labels	-	-
SciTail [307]	NLP	2018	Text Entail.	Prem.-Hyp. → Binary	27K	10.1K / 16.9K
SQuAD 2.0 [541] [540]	NLP	2018	QA	Passage + Q → Ans.	150K+	130.3K / 8.9K
PIQA [46]	NLP	2019	MC QA	Q → Ans.	19K	16K / 3K
Winogrande [562]	NLP	2019	Reasoning	Fill-in-the-blank → Option	44K	-
OSCAR [501]	NLP	2019	Pre-training	Text → Text	50K	-
SuperGLUE [660]	NLP	2019	NLU	Text → Text	-	-
codeSearchNet [261]	Code	2019	-	Code snippets	-	2M
AmazonQA/Products [202]	NLP	2019	QA	Qs → Ans	923K	395M / 20K
CSQA [617]	NLP	2019	QA	Q + Ctx → MC Ans.	12.2K Qs	9.7K / 1.1K
SAMSum [189]	NLP	2019	Dial. Summ.	Dialogues → Summary	16.4K	14.7K / 819
WiC [526]	NLP	2019	Word-in-Context	Text → Binary label	7.5K	5.4K / 1.4K
Hyperpartisan [309]	NLP	2019	Binary Clf.	News → Label	754K	600K / 4K
PAWS [794]	NLP	2019	Para. Ident.	Sent. pairs → Label	108.5K	49.4K / 8K
WoW [138]	NLP	2019	Knowl.-driven Dial. Gen.	Hist. + Topic → Response	-	-
CUB-200-2011 [657]	CV	2011	Fine-grained Classification	Images → Bird Species Labels	11.8K	6K / 5.8K
ELEVATOR [394]	Multimodal	2021	Entity Linking	Text+Image → Entities+Attributes	10K	7K / 1.5K / 1.5K
TB-1k [414]	CV	2021	Medical Classification	Chest X-rays → Labels	1K	800 / 200
BoolQ [106]	NLP	2019	Boolean QA	Question → Yes/No	15.9K	9.4K / 3.3K / 3.2K

Table 5: Comprehensive overview of datasets used for PEFT methods across various domains including NLP, Computer Vision, Multimodal, and Language Models.

Dataset	Domain	Year	Task Type	I/O Format	Size	Train/Test
ro-en	MT	2016	Machine Translation	Romanian → English	614K	610K / 2K / 2K
de-en	MT	2014	Machine Translation	German → English	4.55M	4.54M / 3K / 2.2K
Ha–En	MT	2021	Machine Translation	Hausa → English	50K	45K / 2.5K / 2.5K
(De-En)	MT	2019	Machine Translation	German → English	3.9M	3.9M / 3K / 3K
Fr–De	MT	2020	Machine Translation	French → German	2.8M	2.8M / 3K / 3K
Fr–Es	MT	2018	Machine Translation	French → Spanish	2M	2M / 3K / 3K
News Commentary [632]	MT	2017	Machine Translation	Multilingual → Translations	370K	90% / 5% / 5%
AmazonQA [203]	NLP	2019	Product QA	Product info → Answers	1.4M	80% / 10% / 10%
GPT-3.5 [757]	NLP	2022	Language Generation	Text prompts → Generated text	N/A	Benchmark evals
SRL WSJ & Brown [593]	NLP	2005	Semantic Role Labeling	Text → Argument structures	75K	21/23
aPR-Instruction [363]	Programming	2023	Program Repair	Buggy code + instructions → Fixed code	75K	60K / 7.5K / 7.5K
TinyAgent	NLP	2023	Agent Task Completion	Task descriptions → Actions	10K	8K / 1K / 1K
AddSub [472]	Math	2014	Math Word Problems	Math problems → Numerical answers	395	Cross-validation
SingleEq [321]	Math	2016	Algebraic Problems	Equations → Solutions	508	80% / 20%
CB	NLP	2018	Natural Language Inference	Premise-hypothesis → Labels	250	200 / 50
Flan v2	NLP	-	Instr. Tune	Text → Text	-	-
BBH [611]	NLP	2022	QA	Qs → Ans.	6.5K	-
Flan v2	NLP	-	Instr. Tune	Text → Text	-	-
BBH [611]	NLP	2022	QA	Qs → Ans.	6.5K	-
M2D2 [545]	Multimodel	2022	LM	Text → Tokens	8.5B	-
Alpaca [621]	NLP	2023	Instr. Follow	Text → Text	52K	52K / -
GPT-4 Alpaca [516]	NLP	2023	Instr. Tune	Instr. → Resp.	52K	-
Dolly [36]	NLP	2023	Instr. Follow	Prompt → Resp.	15K	-
Orca [476]	NLP	2023	Reasoning	Prompt → Step-wise	1.6M	-
GPT-4-Turbo [153]	Multi	2023	-	-	128K	-
AI4Bharat Naamapadam [462]	NLP	-	NER	Text → Tags	400K	-
AmericasNLI [154]	NLP	-	NLI	Text → Label	-	-
SIQA [568]	NLP	-	QA	Text → MC Ans.	33.4K	-
TREC [373]	NLP	-	QC	Text → Label	4.5K	5.5K / 500
ScienceQ [691]	NLP	-	Sci-QA	Text → Choice	-	-
Wikitext2 [461]	NLP	-	LM	Text → Tokens	2M	-
Penn Treebank [448]	NLP	-	LM	Text → Tokens	1M	38K / 5.5K
VQA [192] v2.0	Multimodal	2015	Visual QA	Image → Text	265K	-
VisDA-C [519]	Multimodal	2017	Domain Adaptation	RGB Images → Class Labels	280K	152K / 55K
ImageNet-Sketch [672]	Multimodal	2019	Image Classification	Sketch Images → Class Labels	51K	-
DocVQA [451]	Multimodal	2020	Document VQA	Document Images → Text Answer	62.8K	39.5K / 5.2K
FGVC [446]	Multimodal	-	Fine-grained Classification	Aircraft Images → Class Labels	10.2K	3.3K / 3.3K
IWSLT	Multimodal	-	-	- → -	-	-
Caltech101 [167]	Vision	2004	Image Class., Obj. Rec.	Images → Class Labels	9.1K	-
Oxford102 [713]	CV	2008	Fine-grained Cls.	JPEG → 102 lbls	8.2K	7.2K / 1K
CIFAR-10 [331]	CV	2009	Img Cls.	Img → 10 lbls	60K	50K / 10K
CIFAR-100 [331]	CV	2009	Img Cls.	Img → 100/20 lbls	60K	50K / 10K
CIFAR-10-LT [331]	CV	2009	Img Cls. (LT)	Img → 100/20 lbls	60K	50K / 10K
STL-10 [109]	CV	2010	Img Cls.	Img → 10 lbls	13K	5K / 8K
SUN397 [726]	CV	2010	Scene Cls.	Img → 397 lbls	109K	-
SBU [500]	CV	2011	Img Captioning	Img → Caption	1M	-
Stanford Dogs [306]	CV	2011	Fine-grained Cls.	Img → 120 dog breeds	20.6K	14.4K / 6.2K
SVHN [484]	CV	2011	Digit Recog.	Img → Digit lbls	600K	-
OxfordPets [511]	CV	2012	Fine-grained Cls. + Seg.	Img → 37 breeds + masks	7.4K	-
ImageNet1K [559]	CV	2012–17	Img Cls. + Loc.	Img → 1K lbls	1.43M	1.28M / 100K
ImageNet100 [558, 631]	CV	2012	Img Cls.	Img → 100 lbls	132K	127K
Stanford Cars [330]	CV	2013	Fine-grained Cls.	Img → 196 car models	16.2K	8.1K / 8.0K
DTD [102]	CV	2014	Texture Cls.	Img (300x300) → 47 textures	5.6K	40 train / rest test
Food101 [51]	CV	2014	Food Cls.	Img → 101 classes	101K	75.8K / 25.3K
PASCAL Context [475]	CV	2014	Scene analysis	Img → Pixel-wise lbls	10.1K	10.1K
Pascal Context-59 [475]	CV	2014	Seg.	Images → 59 Labels	10.1K	-
Clothing1M [727]	CV	2015	Class.	Images → Noisy	1M	50k / 10k
MICCAI 2015 Abdomen [341]	CV	2015	Seg.	3D CT scans → Organs (13)	30 / 20	-
ModelNet40 [723]	CV	2015	3D Class.	Mesh models → 40 classes	12.3K	9.8K / 2.5K
NABirds [645]	CV	2015	Fine-grained Class.	Bird images → 555 classes	48.6K	23.9K / 24.6K
VQA v2.0 [192]	CV	2017	QA	Image → Text	265K	-
DUT [681]	CV	2017	Object Detection	Images → Binary masks	5.2K	10.6K / 5K
COCO Stuff [59]	CV	2017	Seg.	Images → Stuff + Things	164K	118K / 41K
ADE20K 150 [811]	CV	2017	Seg.	Images → 150 classes	25K	20.2K / 3.4K
ADE20K-847 [811]	CV	2017	Seg.	Images → 847 classes	25K	20.2K / 3.4K
RESISC45 [91]	CV	2017	Scene Class.	256x256 images → 45 classes	31.5K	-

Table 6: Comprehensive overview of datasets used for PEFT methods across various domains, including NLP, Computer Vision, Multimodal, and Language Models.

Dataset	Domain	Year	Task Type	I/O Format	Size	Train/Test
Kinetics-400 [299]	CV	2017	Action Class.	Video clips → 400 labels	650K	-
Something-v2 [191]	CV	2017	Action Class.	Short video clips → 174 labels	221K	169K / 27K
VisDA-C [519]	CV	2017	Domain Adapt.	Images → Domain Class	280K	152K / 55K
EuroSAT [226]	CV	2018	Land Use Class.	Satellite images (64x64) → 10 labels	27K	24.3K / 2.7K
iNaturalist [35]	CV	2018	Species Class.	Wildlife images → 8,142 species	612K	437K / 149K
Taskonomy [776]	CV	2018	Multi-Task Vision	Indoor images → depth, seg., normals	4M	4M
ScanObjectNN [641]	CV	2019	3D Obj. Class.	Point clouds → 15 labels	15K	11.4K
ImageNet-Sketch [673]	CV	2019	Image Class.	Sketches → labels	50.9K	-
ImageNet-A [230]	CV	2019	Image Class.	Real-world images → labels	7.5K	-
VOT22RGBD [558]	CV	2019	RGB-T Tracking	RGB+Thermal → BBoxes	234	-
IMD20 [493]	CV	2020	-	-	35K	-
COD10K [163]	CV	2020	Camouflaged Det.	RGB → Masks	10K	-
CAMO [346]	CV	2020	Camouflaged Seg.	RGB → Binary Masks	1.25K	1K / 250
ORBIT [450]	CV	2021	Few-shot Obj. Recog.	Mobile Vids → Labels	2.7K	70 / 46 users
DepthTrack [429]	CV	2021	RGB-D Tracking	RGB+D → BBoxes	150	80 / 35
IconQA [423]	CV	2021	VQA	Icons → Answers	645K	-
ImageNet-R [228]	CV	2021	Image Class.	Rendered → Labels	30K	-
LasHeR [358]	CV	2021	RGB-T Tracking	RGB+Thermal → BBoxes	730K	-
Pix2Struct [348]	CV	2022	VisLang. Understanding	Screenshots → Text	-	-
VisEvent [689]	CV	2023	RGB+Event Tracking	Text → Tracks	820	500 / 320
Fru92 [241]	CV	2017	Fruit Cls.	Fruit imgs → Labels	69.6K	55.7K / 13.9K
Metaworld [767]	CV	2019	Robot Tasks	Sim states → Actions	481.5K	80 / 20 %
Franka-Kitchen [74]	CV	2020	Robot Control	Kitchen obs. → Actions	566	80 / 20 %
Long-tailed [438]	CV	2019	Imbalanced Cls.	RGB imgs → Labels	100K	80 / 10 / 10 %
Veg200 [241]	CV	2017	Veg Cls.	Veg imgs → Labels	91.1K	72.9K / 18.2K
CottonLeafDisease [45]	CV	2024	Plant Disease Cls.	Cotton imgs → Disease	3K	80 / 20 %
Robust 04 [652]	IR	2004	Info Retrieval	Text docs → Relevance	528K	TREC cross-val
Tufano [239]	IR	2019	Code Repair	Java → Transformed code	58.3K	58.3K / 6.5K / 6.5K
OpenImages V6 [337]	CV	2022	Visual Recognition	RGB imgs → Labels	200K	160K / 40K
Firefly [253]	CV	2022	Text-to-Image	Text → Images	125M	Primarily train
OPUS-100 [492]	Multilingual	2022	Machine Translation	Text → Text	55M	1K/pair
ScanObjectNN [811]	CV	2019	3D Object Classification	Point Clouds → Class Labels	15	2.0K/0.9K
Vizwiz [204]	CV	2018	VQA for the Blind	Img + Q → Ans	31.2K	20.5K/4.3K/6.3K
Flickr30k [437]	CV	2014	Image Captioning	Img → Text	31.8K	29.8K/1K/1K
OKVQA [385]	CV	2019	VQA w/ Knowledge	Img + Q → Ans	14.1K	9K/5K
OCR-VQA [470]	CV	2019	OCR-based VQA	Img + Q → Ans	207.6K	186K/21K
LibriSpeech [506]	Audio	2015	Speech-to-Text	Audio (16kHz) → Text	1,000h	960h + dev/test sets
ImageNet-C [229]	CV	2019	Image Cls	Img → Label	50K	Eval only
DomainNet-126 [518]	CV	2019	Domain Adaptation	Img → Label	586.6K	70%/ 30%
ISTD [678]	CV	2018	Shadow Detection	Img → Clean Img + Mask	1.9K	1.3K/540
PascalCtx-459 [475]	CV	2014	Semantic Seg.	Img → Pixel Labels	10.1K	5K/5.1K
CRer	CV	2021	Crop Disease Detect.	Img → Label/Mask	3K	80%/ 20%
CASIA [759]	CV	2010–18	Face Recog.	Face Img → ID	494.4K	90%/10%
IMD20 [439]	CV	2020	Face Recog.	Face Img → ID	34K	70%/30%
CUHK [371]	CV	2012	Person Re-ID	Img → ID	14.1K	1.4K/100 IDs
CHAMELEON [249]	CV	2018	Camouflaged Obj Det.	Img → Mask	76	Test only
Kinetics-700 [63]	CV	2020	Action Recog.	Video → Action	650K	600K/50K
DF-20M mini[525]	CV	2022	DeepFake Detect.	Face Img/Vid → Real/Fake	2–3M	80%/20%
RIGA+	Medical	2019	Retinal Segmentation	Fundus Img → Mask	750	650/100
SCGM[487]	Medical	2017	Spinal Cord Seg.	MRI → Mask	80	60/20

Table 7: Comprehensive overview of datasets used for PEFT methods across various domains including NLP, Computer Vision, Multimodal, and Language Models.

Year	Model	Application	PEFT Method	Datasets
2025 [630]	Mistral 7B [274]	Code conversion system	LoRA	APPS, Conala, CodeAlpacaPy
2025 [185]	BERT [136], mT5 [537], mGPT [5]	Low-resource sentiment analysis, NER, QA	Full-Finetune, LoRA, AdaLoRA, DoRA	Persian NLP
2025 [273]	GLM4-9B [150]	Language understanding, sentiment analysis	MoE-LoRA, X-LoRA, TAAS-Net, IA^3, LoRA	GLUE, IMDB, Agnews
2025 [178]	LLaMA [633], Gemma [626], GPT-4o [5]	Immigration law and insurance	LoRA, QLoRA, DoRA, Prompt Tuning	InsuranceQA, USCIS data
2025 [13]	DistilBERT [567], RoBERTa [413], LLaMa-7B [633]	Common sense reasoning	LoRAShear, LLM-Pruner, LoRAPrune, LoRA variants	BoolQ, PIQA, HellaSwag
2024 [124]	LLaMA [633], RoBERTa [413], TinyLlama [787]	Token prediction, pre-training	Full-Rank, GaLore variants, LoRA, ReLoRA	C4, GLUE, TinyAgent
2024 [590]	Llama-2 [633], GIZA++ [496], xlm-roberta [115]	Cross-lingual transfer	Handholding ICL/PEFT variants	Amazon Massive
2024 [212]	LLaMA-7B [633], Mistral-7B [274]	API prediction, QA tasks	LoRA, AdapterH, MEFT, FT	NaturalQuestion, SQuAD
2024 [618]	DeBERTaV3, Llama/2 [633], GPT-3.5 [408]	NLU, generation, reasoning	GloC, SoRA, AdaLoRA, Adapter variants	GLUE benchmarks
2024 [1]	Vicuna-7b/13b [807]	Text classification	PEFT+ICL, ICL, LoRA, 0-shot	SST2, TREC, AG News
2024 [744]	BLOOM [428], TigerBot [86], Mistral [274]	Cross-lingual vocabulary	LoRA with LAPT	OSCAR, CC-100
2024 [597]	ChatGLM-6B [627]	Medical NER, text tasks	CGC-LoRA, LoRAHub, MOE-LoRA	PromptCBLUE, Firefly
2024 [48]	LLaMA-1/2/3 [152, 633], Mistral [274]	Quantized language generation	LR-QAT, QAT, LoRA, PEFT	SlimPajama, WikiText-2
2024 [208]	Mamba, Mamba-2 [194]	Natural language tasks	MPFT+LoRA, MPFT	MMLU
2024 [520]	Llama-2 [633]	Text Reranking	Q-PEFT variants (MSS, B25, Contriever)	WebQ, TriviaQA
2024 [458]	Mistral-7B [274], ChatGPT-4 [5]	Log parsing	LoRA, PEFT	LogPai
2024 [165]	BERT [136], GPT variants [808, 408]	FL finetuning in edge servers	FedPipe, LoRA, FedAdapter	20NEWS, E2E
2024 [388]	GPT variants [408, 808]	Split learning LLM finetuning	FedLoRA, CenLoRA, SplitLoRA	E2E
2024 [325]	BERT/LLaMa variants [413, 633]	Language understanding, QA	LoRA, Adapter, tuning methods	GLUE, SQuAD, BoolQ
2024 [266]	Mistral-7B [274], Llama-2-7B [633]	LLM personality manipulation	PEFT, IKE	PersonalityEdit
2024 [312]	GPT-Neo [182], GPT-J [100], LLaMA [633]	Task-specific adaptation	PEQA, PEFT+PTQ variants, LoRA	Wikitext2, Alpaca
2024 [738]	Llama-2-7b [633]	Translation, summarization	NEFT, FT variants, LoRA	News Commentary
2024 [716]	RoBERTa [413]	Classification, NER	FT, Adapter, WARP, InfoPrompt	CoLA, SST-2, ACE2005
2024 [408]	BERT [136], GPT2 [408]	Language understanding	CLS-FT, PET-FT, P-tuning	LAMA, SuperGLUE
2024 [164]	InCoder [173], CodeGen [489], Llama2 [633], Code Llama [554], CodeT5 [695], CCT5 [381]	Code Change Tasks	LoRA, Prefix-tuning	MCMD, CodeSearchNet
2024 [209]	Openchat8B [667]	Single and Multi-Downstream Tasks	DoRA, LoRAMoE, LoRA, MixLoRA, MixLoRA-Dy, MoLA, SLIM	OBQA, SIQA, BOOLQA, CSQA, HellaSwag, WinoGrande, ARC-e, ARC-c, MMLU, GSM8K, PIQA
2024 [443]	T5 [537], BART [357], OPT [791], BLOOM [428], Llama 2 [633]	Data to text generation	QLoRA, Full FT	E2E, ViGGo, WikiTableText, DART, WebNLG
2024 [19]	Llama2-7b, Llama2-3b [633]	Factuality, reasoning, multilinguality, coding tasks	SpIEL-MA, SpIEL-AG, IA^3, LoRA, Full FT	Flan v2, GPT4-Alpaca, Tulu v2
2024 [786]	DialoGPT [793], RoBERTa [413], LLaMA2-7B [633], LlaMA2-13B [633]	Personalized response generation	LoRA	AmazonQA/Products, Reddit, MedicalDialogue
2024 [362]	CodeLlama-7B [554], DeepSeek-Coder-Base-6.7B [198] [43], CURE [197], RewardRepair [756], Recorder [821], INCODER-1B, INCODER-6B [173]	Automated Program Repair	FMFT, IA^3, LoRA, p-tuning, Prefix-tuning	APR-Instruction dataset
2024 [647]	T5-BASE [537], GEMMA-2B [626]	Natural language tasks	CLAM-3, CLAM-2, VeRA, LoHA, LoRA, IA^3, QIA^3, CLAM-Q→3, CLAM-PQ→3, QVeRA, QLoHA, QLoRA, PLoRA	GLUE, SuperGLUE
2024 [762]	Gemma-7B [626], Llama 2-7B, Llama 2-13B [633]	Question answering (QA), multi-hop reasoning, counterfactual reasoning	LOFIT, ITI, RepE, 0-shot	TruthfulQA, MQuAKE, CLUTRR

Table 8: Parameter-Efficient Fine-Tuning (PEFT) Methods in NLP Models (2024-2025)

Year	Model	Application	PEFT Method	Datasets
2024 [732]	LlaMA-2 7B [633], LlaMA-2 13B [633], Gemma 2B [626]	Question-answering, sentence level tasks, instruction tuning	PEDRO, BitFit, IA^3, SSP, AdaLoRA, LoRA, Learned-Adapter, Housbly-Adapter, LPT, P-tuning v2, Full-FT	SQuAD, SuperGLUE (BoolQ, COPA, and ReCoRD), GLUE (SST-2, RTE, QNLI), Alpaca dataset, MT-Bench, MMLU, BBH
2024 [464]	OPT-13B [791], LLaMA30B, LLaMA-2-70B [633]	Co-serving system for token generation	HuggingFace (HF) PEFT, S-LoRA, LoRA, IA^3, Adapter	Chatbot instruction prompts, ChatGPT Prompts, WebQA, Alpaca, PIQA
2024 [557]	$RoBERTa_{Large}$ [413], $BERT_{BASE}$ [136] [324], $T5_{SMALL}$, $T5_{BASE}$ [537], Flan-T5-base, Flan-T5-xl [101], OPT [791], GPT2 [408] [808] [408], ViT [146]	Sentiment classification, paraphrase detection, natural language inference, linguistic acceptability, semantic textual similarity, question answering	DEFT, ADA-DEFT, PEFT, ADA-PEFT, LoRA, Adapter, Prefix-Tuning, Prompt-Tuning	MNLI, QQP, QNLI, SST-2, STS-B, MRPC, RTE, SQuAD
2024 [181]	OPT6.7B [791], BLOOM-7B [428], LLaMA-7B [633]	QA, multichoice science questions, problem compilation and concluding tasks	DLoRA, FT	OBQA, PIQA, SIQA, Winogrande, BoolQ, HellaSwag, ARC-easy, ARC-challenge
2024 [584]	Galactica [625], Text+Chem T5 [97], LlaSMol [764], LLaMA [633], Tag-LLaMA [633]	Language domain, protein sequences and SMILES molecule representations, drug discovery	Linear Probing, Prompt Tuning, LoRA, TagLLM	OPUS-100, Flores-101, binding_affinity, Therapeutics Data Commons benchmark, SMILES, TDC benchmark
2024 [195]	RoBERTa-Large [413], OPT-1.3B, OPT-6.7B [791]	Natural language understanding, QA tasks	LoRA, Full-FT, Adapter, LayerDrop, Offsite-Tuning, LLM-Pruner	GLUE, SuperGLUE, OpenBookQA, PIQA, ARC-Easy, ARC-Challenge, SciQ, WebQuestions
2024 [651]	Llama 2 Chat 7B, Llama 2 Chat 13B, Llama 2 Chat 70B [633]	Proprietary documents and code repositories preparation	LoRA, QLoRA	Common crawl, The Pile, Dolly, Orca, Alpaca, Vicuna
2024 [364]	MISTRAL-7B [274], DEEPSEEK-CODER-BASE-6.7B [43] [198], LLAMA3-8B [152], LLAMA2-7B, LLAMA2-13B [152] [633]	Question classification	GMT, One-off Drop, SFT, HFT, Random Mask	Magicoder-Evol-Instruct-110K, MetaMathQA, GSM8k, MATH, TÜLU V2
2024 [753]	LLaMA [633], GPT-J [100] [408], BLOOMz [479], LLaMA3 [152]	Common sense reasoning	LoRA, LoRA + IST, Full Fine-tuning, Series Adapter, Parallel Adapter	BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, OBQA, GSM8K, AQuA, MAWPS, SVAMP
2024 [624]	TinyBERT [277], MobileBERT [607], DistilBERT [567], standard BERT [136], Llama-2-7b [633]	Clinical decision-making tasks	LoRA, IA3, Full FT	MIMIC-III MP, MIMIC-III LoS, MIMIC-III ICD-9 Triage, I2B2 2010 RE,2010,2012,2014
2024 [743]	LLaMA-2-13B [633], BLOOM-7B [428]	Token generation	LoRA, Soft Prompt	Wikitext2
2024 [327]	RoBERTa-base, RoBERTa-large [413], DeBERTaV3-base [220], BART-large [357], $BLOOM_z$ [479], LLaMA7B [633], LLaMA13B [633], GPT-J6B [100] [408]	Common sense reasoning and mathematical reasoning	LoRA, AdaLoRA, Prefix, Propulsion, Bitfit, $(IA)^3$, Prompt Tuning	GLUE, SQuAD, XSum, CNN/DailyMail
2024 [564]	FlanT5-XXL [101]	Text classification and generation for privacy-preserving personalization of LLM	PEFT, RAG, PEFT+RAG	LaMP benchmark
2024 [375]	Llama-2 [633], Falcon [14]	Robust emotion recognition	LoRA	IEMOCAP
2024 [502]	T5-Base [537], BART-Large [357], T5-Large	Sentiment analysis, similarity and paraphrase, natural language inference tasks for resource-constrained edge devices	LoRA, Adapter, Full FT	GLUE
2024 [181]	OPT6.7B [791], BLOOM-7B [428], LLaMA-7B [633]	QA, problem compilation and concluding tasks, multi-choice science questions on edge devices	DLoRA	OBQA, PIQA, SIQA, Winogrande, HellaSwag, ARC-easy, ARC-challenge
2024 [603]	LLama2-7B [633]	Generative text using concurrent LLM serving	LoRA, BOFT, S-LoRA, SpartanServe	Personalized text data of varying lengths
2024 [586]	Baichuan2-13B [747]	Domain-specific Chinese relation extraction	LoRA	FinRE, SanWen

Table 9: Parameter-Efficient Fine-Tuning (PEFT) Methods in NLP Models (2024-2025)

Year	Model	Application	PEFT Method	Datasets
2023 [683]	T5-Base, T5-Large, T5-XL [537]	Natural language understanding tasks	FT, PT, Prompt Tuning, XPrompt, ResPrompt, APROMPT	SuperGLUE
2023 [78]	BERT [324], RoBERTa [413], ALBERTAxxlarge-v2 [560]	Natural language understanding tasks in few shot and fully supervised settings	P-Tuning, P-Tuning v2, PTP+RM, PTP+RG, PTP+A2T, PTP+PGD, PET Best	SuperGLUE, FewGLUE
2023 [94]	T5 [537]	Natural language understanding	Full FT, Prompt Tuning, P-Tuning, SMoP	SuperGLUE
2023 [587]	T5-Base [537], CLIP-T5[535]	Natural language processing (NLP), visual question-answering task, image caption generation	LoRA, FT, Adapters, BitFit, PT, SPoT, ATTEMPT, MPT, DEPT	GLUE, SuperGLUE, WinoGrande, Yelp-2, SciTail, PAWS-Wiki, VQA, MSCOCO
2023 [773]	T5 v1.1 + LM adaptation	Instruction tuning	T0, $(IA)^3$, LoRA, MoV, MoLoRA	ANLI, CB, RTE, WSC, WiC, COPA, WNG, HS
2023 [426]	GPT-2, GPT-3, GPT-3.5, GPT-4 [536][408][5], DialoGPT [793], GODEL [515], Blenderbot [589], ChatGPT, DOHA [527], T5 [537]	Toxicity reduction, lexically constrained generation, open-ended generation, dialogue safety control, and knowledge grounded dialogue	IPA, IPA*, PPLM, GeDi, DEXPERTS, DAPT, PPO, QUARK	RealToxicityPrompts, XSum, CommonGen, ASAFETY, WoW
2023 [343]	RoBERTa-Large [413]	Natural language understanding	LoRA, FFT, BitFit, Adapters, MaM, S-MaM, U-MaM, WARP, S-BitFit, U-BitFit	GLUE
2023 [378]	RoBERTa-Large [413]	Natural language understanding and translation tasks	LoRA, Full FT, Linear FT, Linear FT_{norm}, Diff Pruning, FISH Mask, Adapter, Pfeiffer Adapter, BitFit, Prefix Tuning, MAM Adapter, PaFi	GLUE, VariousGLUE
2023 [125]	T5-Large [537]	Sequence labeling, relation extraction and joint entity relation extraction task	TANL, ASP, Fixed FISH, FISH-DIP, DygiePP	CoNLL'03, OntoNotes, CoNLL'04, ACE2005, MultiWoz 2.1, CoNLL'05 SRL WSJ and Brown
2023 [531]	FLAN-T5	Generation and classification tasks	LoRA, Full FT, IA_3, BitFit, Prompt Tuning	AG News, CoLA, E2E, NLG, SAMSum
2023 [351]	T5 [537]	Classification, question answering, summarization and speech recognition	Parallel Adapter, CoDA, Prefix Tuning, Sequential Adapter	C4, LibriLight, Pix2Struct, OCR-VQA, DocVQA, Screen2Words, MNLI, RTE, BoolQ, SQuAD, XSum, LibriSpeech
2023 [98]	GPT-2	Language modeling	Single Adapter, AdapterSoup, Oracle, Hierarchy adapter	M2D2, C4
2023 [224]	BERT-Base, RoBERTa-Base	Sentiment analysis, question-answering, natural language inference, and commonsensereasoning	FT, AdapterFusion, Adapter, MerA	MRPC, SST-2, MNLI
2023 [365]	Longformer-base, RoBERTa-base	Long sequence language tasks	Prefix-Propagation, FT, PT, Prefix-tuning	ArXiv, 20-newsgroups, Hyperpartisan, WikiHop
2023 [796]	BERT-base, BERT-large, RoBERTa-large, DeBERTa-xlarge	Natural language understanding and named entity recognition tasks	FT, PT, APT	BoolQ, COPA, RTE, WiC, WSC, CoNLL03, CoNLL04, OntoNotes
2023 [177]	jiant [661]	Natural language understanding	LoRA, Random, Mixout, BitFit, MagPruning, Adapter, DiffPruning, ChildPruning, SAM	GLUE, SuperGLUE
2023 [788]	DeBERTaV3-base [220], BART-large [357]	Natural language understanding, question answering, natural language generation	LoRA, AdaLoRA, Full FT, BitFit, HAdapter, PAdapter	GLUE, SQuADv1, SQuADv2, XSum, CNN/DailyMail
2023 [139]	DeBERTaV3-base [220], RoBERTa-large [413]	Natural language understanding	LoRA, SoRA	GLUE
2023 [746]	LlaMA2-7B [633]	Commonsense reasoning and out of distribution	MAP, MC Drop, Ckpt Ens, Ensemble, LLLA, LA	Winogrande-small (WG-S), Winogrande-medium (WG-M), ARC-Challenge (ARC-C), ARC-Easy (ARC-E), openbook QA (OBQA), BoolQ, MMLU Big-Bench Hard (BBH)
2023 [255]	FLAN-T5 [101]	Multiple-choice questions from a variety of domains	LoRA, FFT, IA3	
2023 [401]	ChatGPT [5], Huatuo [671]	LLM-driven medical applications and LoRA	LoRA, P-Tuning, Task-Arithmetic, LoRAHub, MOELoRA	PromptCBLUE
2023 [75]	T5-base/3b [537], RoBERTa [413], BART [357]	Natural language understanding and generation tasks	LoRA, Adapter, BitFit, Prefix, PA, S_4	GLUE, XSum, WMT 2016 en-ro

Table 10: Parameter-Efficient Fine-Tuning (PEFT) Methods in NLP Models (2023)

Year	Model	Application	PEFT Method	Datasets
2023 [813]	BERTbase [324]	Text classification tasks	LoRA, FFT, Prefix, AdaMix, Serial, UniPELT, Parallel, MAM, AutoPEFT	GLUE, SuperGLUE
2023 [254]	LLaMA [633], BLOOMz [479], GPT-J [100]	Arithmetic reasoning and commonsense reasoning	LoRA, Prefix, Series, Parallel	MultiArith, GSM8K, AddSub, AQuA, SingleEq, SVAMP, BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, OBQA
2023 [800]	T5-Base, T5-Small, T5-Large [537]	Multi-task learning for natural language understanding	Adapter, FT, PT, SPoT, HF, ATP, HD, MPT, PHA	CoLA, SST-2, STS-B, MRPC, QQP, MNLI, QNLI, RTE, SciTail, BoolQ, WiC, CB, WSC
2023 [741]	ALBERT [340], DistilBERT-base [567], BERT-base [136], RoBERTa-large [413], LLaMA-7B (INT4) [633]	News classification, topic classification, QA using federated fine-tuning of LLMs	LoRA, Adapter, BitFit, Full-FT	AGNEWS, YAHOO, YELP-Polarity, SQuAD-v1.1
2022 [434]	T5-Large, T5-XL, and T5-XXL [537]	Natural language understanding	Fine-Tuning, P-Tuning, Prefix-Tuning, Prompt-Tuning, XPROMPT	SuperGLUE
2022 [724]	RoBERTa [413] and EFL [687]	Natural language understanding	Transformer fine-tuning, Prompt tuning, Adapter, S-IDPG-PHM, S-IDPG-DNN, M-IDPG-PHM-GloV, M-IDPG-PHM, M-IDPG-DNN, Compacter, P-Tuningv2	MPQA, Subj, CR, MR, SST-2, QNLI, RTE, MRPC, STS-B, QQP
2022 [406]	RoBERTaLARGE [413], DeBERTaLARGE [221], GPT2LARGE [408]	Single-sentence and sentence-pair classification and PTM tasks	LoRA, Adapter, AdapterDrop, BitFit, Prompt Tuning, P-tuning v2, S-IDPG-PHM, LPT w/ NPG, LPT w/ MPPG, LPT w/ APPG, LPT w/o PG	SST-2, MPQA, MR, Subj, TREC, MNLI, MRPC, QNLI, QQP, RTE
2022 [251]	T5-large [537]	Natural language understanding	LoRA, Fine-tune, BitFit, Low Rank Adapter, Adapter, LNFit, S^3PET	GLUE, SUPERGLUE
2022 [693]	RoBERTa-large [413], BERT-base [136]	Supervised and few-shot NLU and NLG tasks	Full Fine-tuning, Pfeiffer Adapter, Houlsby Adapter, LoRA, AdaMix Adapter, BitFit, Prefix-tuning, UNIPELT, Lin Adapter, AdaMix LoRA	MNLI, QNLI, SST2, QQP, MRPC, CoLA, RTE, STS-B
2022 [395]	T0, T5 [537], GPT-3 [408]	Few-shot classification tasks	T-Few, PET, BitFit, Human baseline	RAFT
2022 [288]	TwinBERT [425], ColBERT [305]	Document Reranking	Full FT, LFT, Prompt-tuning, Prefix-tuning, LoRA, LoRA+, Prefix-tuning $\rightarrow$ LoRA, LoRA $\rightarrow$ Prefix-tuning, SS Prefix-tuning, SS LoRA	Robust04, ClueWeb09b, MS-MARCO
2022 [644]	RoBERTa [413], GPT-Medium	Natural language understanding and language generation	LoRA, Fine Tune, FLOP, DyLoRA	GLUE, E2E, DART, WebNLG
2022 [175]	BERT-base [136], BERT-large [136], RoBERTa-base, RoBERTa-large [413]	Added token-dependent biases to the shifts by proposing AdapterBias for PLMs	Full-FT, Adapters, Diff-pruning, BitFit, LoRA	GLUE
2021 [447]	BERT-base [136], BART-large [357]	Natural language understanding	LoRA, FT, BitFit, Adapter, Prefix-tuning, UNIPELT	GLUE
2021 [372]	GPT-2 [408], BART [357]	Natural language generation tasks	FT-TOP2, FINE-TUNE, ADAPTER, PREFIX, SOTA	E2E, WebNLG, DART
2021 [407]	BERT-large [136], RoBERTa-large [413], GLMxlarge, GLMxxlarge [150]	Model scaling and NLU tasks	FT, PT, PT-2, MPT-2	BoolQ, CB, COPA, MultiRC, ReCoRD, RTE, WiC, WSC, CoNLL03, OntoNotes 5.0, CoNLL04, SQuAD 1.1 dev, SQuAD 2.0 dev, CoNLL12, CoNLL05, WSJ, CoNLL05 Brown
2021 [354]	T5 [537], GPT-3 [408]	Question answering (QA) and paraphrase detection in zero-shot settings	Prefix tuning, Prompt tuning	SuperGLUE
2021 [653]	T5 [537]	Natural language understanding tasks using single and multitask training	PROMPTTUNING, MODELTUNING, SPoT, MULTI-TASKMODELTUNING	GLUE, SUPERGLUE
2021 [604]	RoBERTa [413], T5 [537]	Sentiment analysis, natural language inference, ethical judgment, paraphrase identification, QA, Summarization	Prompt tuning, TPT, Random Prompt, Distance Minimizing, Task Tuning	IMDB, SST-2, laptop, restaurant, Movie, Tweet, MNLI, QNLI, SNLI, deontology, justice, QQP, MRPC, SQuADNQ-Open, Multi-News, SAMSum
2021 [532]	$\text{BART}_{\text{BASE}}$, $\text{BART}_{\text{LARGE}}$ [357]	Text classification, question answering, conditional generation, etc.	Fine-tuning, Prompt Tuning, IPT	CrossFit Gym
2021 [610]	$\text{BERT}_{\text{LARGE}}$ [136], ResNet-34	Natural language understanding, distributed training, efficient checkpointing	Dense Fine-tuning, Random Mask, Bit-Fit, Diff Pruning, FISH Mask	GLUE, CIFAR-10

Table 11: Parameter-Efficient Fine-Tuning (PEFT) Methods in NLP Models (2021-2023)

Year	Model	Application	PEFT Method	Datasets
2021 [18]	mBERT [737] [136], XLM-R [115]	Zero-shot cross-lingual transfer	LT-SFT, RAND-SFT, MAD-X, BITFIT, LT-SFT TA-ONLY, MAD-X TA-ONLY	Universal Dependencies 2.7, MasakhaNER, CoNLL 2003, AmericasNLI, MultiNLI
2021 [442]	T5 [537]	Natural language understanding	HYPERFORMER, Adapters	GLUE
2021 [742]	BERTLARGE [136], XLNetLARGE, RoBERTaLARGE [413], ELECTRALARGE [104]	Single-sentence and a pair of sentence classification tasks	Vanilla Fine-tuning, CHILD-TUNING, Weight Decay, Top-K Tuning, Mixout, RecAdam, R3F	GLUE
2021 [775]	BERT [136], RoBERTa [413]	Sentence level and token level NLP tasks	Full-FT, BitFit	GLUE, PTB POS-tagging
2021 [186]	Transformer [646]	Machine translation	scratch (100%), src,tgt (8%), src,tgt+body (75%), src,tgt+xattn (17%), src,tgt+randxattn (17%)	Ro–En, Ja–En, De–En, Ha–En, Fr–Es, Fr–De
2021 [247]	GPT-2, GPT-3 [408]	Natural language to SQL queries, natural language inference, conversation summarization, natural language generation tasks	LoRA, FT, BitFit, Adapter, Prefix-layer tuning, Prefix-embedding tuning	E2E, WikiSQL, MultiNLI, SAMSum
2021 [227]	T5 [537]	Natural language understanding tasks	ADAPTER, PFEIFFER-ADAPTER, ADAPTERDROP, ADAPTER-LOWRANK, PROMPT TUNING, INTRINSIC-SAID, BITFIT, PHM-ADAPTER, COMPACTER	GLUE, SuperGLUE
2021 [822]	BLEU [510]	Multilingual machine translation	Bilingual, Multilingual, Knowledge Distillation, Serial, CIAT-basic, CIAT-block, CIAT	IWSLT^2, OPUS-100, WMT
2021 [217]	BARTLARGE [357], mBARTLARGE [410], RoBERTaBASE [413]	Machine translation, text summarization, language understanding, text classification	LoRA, Parallel adapter, Full-FT, Bitfit, Adapter, Prefix, PA, MAM adapter, Pfeiffer adapter, Prompt tuning, Prefix tuning	XSum, WMT 2016 en-ro, MNLI, SST2
2020 [522]	BERT-baseuncased [136], RoBERTa-base [413]	Commonsense reasoning, sentiment analysis, natural language inference, sentence relatedness	AdapterFusion, Single-Task Adapters, Multi-Task Adapters, Fine Tuning, Fusion w/ ST-A, Fusion w/ MT-A	Hellaswag, Winogrande, CosmosQA, CSQA, SocialIQA, IMDb, SST, MNLI, SciTail, SICK, RTE, CB, MRPC, QQP5, Argument, BoolQ
2020 [199]	BERTLARGE [136], RoBERTa [413], XLNet [752]	Natural language understanding, question answering	LoRA, Full finetuning, Adapters, Non-adaptive diff pruning, Diff pruning	GLUE benchmark, SQuAD
2020 [7]	RoBERTa [413], BERT [136], BART [357], Electra, Alberta [340], XLNet, T5 [537], XLM-R [115]	Sentence prediction tasks	Fine Tuning, SAID, DID	MRPC, QQP
2019 [244]	BERT [136]	Text classification tasks	Adapters	GLUE, SQuAD

Table 12: Parameter-Efficient Fine-Tuning (PEFT) Methods in NLP Models (2019-2021)

Year	Model	Application	PEFT Method	Datasets
2025 [387]	FaceT-B, PLFace	Face recognition	DPEFT	MS1MV3, CASIA-WebFace-masked
2024 [279]	ViT-B/16 [146]	Image classification	VPT-Deep, Full FT, Linear probing, VPT-Shallow, Adapter, AdaptFormer, LoRA, NOAH, Convpass$_{share}$, Convpass$_{attn}$, Convpass	VTAB-1K
2024 [761]	Swin-L, Swin-B, Swin-L [415]+Cascade Mask RCNN	Dense predictions in computer vision	Full FT, Fixed FT, Bitfit, Norm-Tuning, Partial-1, Adapter, LoRA, AdaptFormer, LoRand, E_3VA, E_3VA+, E_3VA++	MS COCO, PASCAL VOC and ADE20K
2024 [67]	ViT-B [146] SAM [315]	Medical image segmentation	Ladder-Side Tuning	MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge
2024 [733]	Segformer [728], Swin Transformer [415]	Multitask dense scene understanding	Single-task Full Fine-tuning, Fine-tuning Decoders, Multi-task Full Fine-tuning, Multiple Bitfit, Multiple Relative bias, Multiple LoRA, Multiple Adapter, Multiple Low-rank adapter, Shared BitFit, Shared Relativebias, Shared LoRA, Shared Adapter, Shared Low-rank adapter, Hyperformer, Polyhistor, Polyhistor-lite, VMT-Adapter, VMT-Adapter-Lite	PASCAL Context
2024 [670]	ResNet-50, ResNet-101 [219], ViT [146], Swin Transformer [415]	Image classification, long-tail distribution, few-shot learning	Retraining, Head-tuning, FT, Adapter, Bias, VPT, LION	CIFAR10, CIFAR100, ImageNet100, Flower, Stanford Dogs, Stanford Cars, Clothing
2024 [176]	ViT-B/16 [146], Swin-B [415]	Visual recognition, object counting or depth prediction, domain generalization	BitFit, VPT, LST, AdaptFormer, LoRA, NOHA, FacT, SSF, DTL, DTL+, DTL+*	VTAB-1K, Aircraft, Pets, Food-101, Cars and Flowers102
2024 [34]	ViT-S(DINO) [146], ViT-S(DeiT) [146]	Few-shot image classification	Full FT, Bias, Adapter, LoRA, Ladder, Prompt-Shallow, Prompt-Deep, eTT, LN-TUNE, ATTNSCALELITE, ATTNSCALE	META-DATASET and ORBIT
2024 [620]	Point-BERT [136], Point-MAE [31], PointM2AE [31]	3D shape classification	Full Fine-Tuning, Prompt Tuning, Adapter Tuning, LoRA, Bias Tuning, Point-PEFT	ScanObjectNN, ModelNet40
2024 [592]	ESM2	PPI prediction, multiplicity and symmetry prediction task	FT, PEFT, LoRA	PPI data, Homooligomer Symmetry data
2024 [816]	LLaVA-1.5(7B, 13B) [396], ShareGPTv4(7B) [79], Qwen-VL-Chat(7B) [29]	Visual question answering, visual reasoning, image caption using multimodal LLM	Adapter, LoRA, IA3, Prefix	ScienceQA, Vizwiz, IconQA, Flickr30k, OKVQA, OCRVQA, VQAv2
2023 [639]	ViT-B/16, self-supervised (MAE), CLIP [535]	Visual recognition tasks	Linear-probing, Fine-tuning, VPT, VQT	VTAB-1k
2023 [740]	CLIP ViT-L/14	Open vocabulary semantic segmentation	SimSeg, OvSeg, MaskCLIP, SAN	COCO Stuff, Pascal VOC, Pascal Context-59, Pascal Context-459, ADE20K150, ADE20K-847
2023 [729]	DiT	Class-conditioned image generation	Full Fine-tuning, Adapter, BitFit, Visual Prompt Tuning (VPT), LoRA, DiffFit	Food101, SUN397, DF-20M mini, Caltech101, CUB-200-2011, ArtBench-10, Oxford Flowers, Stanford Cars
2023 [225]	ViT-B-224/32 via unsupervised pretraining (CLIP), Supervised ViT	Image classification	Fine-tuning	ImageNet-1k
2023 [72]	Vision transformers, hierarchical swin transformers [415]	Visual classification task	BitFit, VPT-Shallow, VPT-Deep, Adapter, AdapterFormer, LoRA, NOAH, Full FT, Linear probing, SSF, FacT-TT, FacT-TK, EFFT	VTAB-1K
2023 [583]	ViTs [146], NFNets [52], and ResNets [219]	Manipulation task	Full FT, Adapters, Pretrained Feat. (DR)	Metaworld, Franka-Kitchen, RGB-Stacking task suites
2023 [422]	LLaMA-Reviewer, CodeReviewer and AUGER	Automating code review tasks	Prefix tuning, LoRA	CRer, Tufano
2023 [405]	SETR [347]	Detection tasks	Full-tuning, OnlyDecoder, VPT, AdaptFormer, EVP	CASIA, IMD20, SBU, ISTD, CUHK, DUT, COD10K, CHAMELEON, CAMO
2023 [430]	ViT-L/14, ViT-B/16 [146]	Image and video classification, semantic segmentation	VPT, Adapter, AdapterFormer, LoRA, NOAH, Full FT, Linear probing, SSF, ReAdapter	VTB-1k, Something-Something V2, ADE20K
2023 [275]	ViT-B/16 [146]	Image and visual classification	BitFit, Prefix tuning, VPT, Adapter, AdapterFormer, Full FT, Linear probing, U-Tuning	CIFAR-100, FGVC

Table 13: Parameter-Efficient Fine-Tuning (PEFT) Methods in Vision Models (2023-2024)

Year	Model	Application	PEFT Method	Datasets
2023 [216]	ViT-B/16 [146]	Recognition tasks	MLP-3, PROMPT-SHALLOW, PROMPT-DEEP, ADAPTER, ADAPTFORMER, NOAH, SPT-ADAPTER, LINEAR, PARTIAL-1, BIAS, LORA, SPT-LORA, SPT-ADAPTER (FULL), VPT	CUB-200-2011, NABirds, Oxford Flowers, Stanford Cars, Stanford Dogs, VTAB-1k
2023 [71]	ResNet-18, ResNet-50 [219]	Transfer learning tasks	RLMVP, FLM-VP, ILM-VP	Flowers102, DTD, UCF101, Food101, GTSRB, SVHN, EuroSAT, OxfordPets, StanfordCars, SUN397, CIFAR10/100, ABIDE
2023 [183]	ViT-B/16 [146]	Continual learning tasks	Sequential finetuning, LAE, Joint-FT, Prompt Tuning, Prefix Tuning, LoRA, Adapter	CIFAR100, ImageNet-R
2023 [278]	ViT-B/16 [146]	Classification tasks	BitFit, VPT-Shallow, VPT-Deep, Adapter, AdapterFormer, LoRA, NOAH, Full FT, Linear probing, FacT-TT, FacT-TK	VTAB-1K
2023 [780]	Point-BERT [136], Point-MAE [31], ACT [145]	Object classification, few-shot learning, part segmentation	Full FT, IDPT, VPT	ScanObjectNN, ModelNet40
2023 [820]	RGB-based foundation tracking model	Multimodal tracking	Full FT, ViPT, Prompt-shaw, Prompt-deep, ViT-shaw	Depthtrack, VOT22RGBD, RGBT234, LasHeR, VisEvent
2022 [90]	ViT [146]	Object detection, instance segmentation, and semantic segmentation	PVT-Tiny, PVTv2-B, ViT, ViTDet, ViT-Adapter	COCO, ADE20K
2022 [193]	ResNet18, ResNet50 and ResNet152 [219], VGG16 and VGG19, SELDnet	Image classification and sound event detection tasks	PHResNet, PHVGG16, PHSELDnet	SVHN, CIFAR10, CIFAR100, ImageNet, L3DAS21 challenge Task 2
2022 [376]	ViT-B/16 [146], Swin Transformer [415], ConvNeXt-B [416], AS-MLP-B [194]	Image classification tasks	Full fine-tuning, linear probing, Adapter, Bias, VPT, SSF	FGVC, VTAB-1k, CIFAR-100, ImageNet-1K
2022 [141]	ViT [146]	Long-tailed image classification tasks	Linear Probe, Full fine-tune, LPT	CIFAR100-LT, Places-LT, iNaturalist2018
2022 [276]	Resnet-50 [219], SimCLR [82]	General classification, fine-grain classification tasks	BiT, TUP, SimCLR-LP, DnA, DnA-MoCo	iNaturalist, CIFAR100, EuroSAT, Food101
2022 [252]	Ushape [161] segmentation model	Medical image segmentation	Intra-Domain, DA, Self-Training, BEAL, DoCR, U-D4R, FSM, ProSFDA	RIGA+, SCGM
2022 [81]	ViT-B/16 [146], ImageNet-21k, MAE, VideoMAE	Image and video recognition tasks	Full FT, Linear probing, VPT (Visual Prompt Tuning), AdaptFormer-1, AdaptFormer-4, AdaptFormer-64	CIFAR-100, Street View House Numbers (SVHN), Food-101, SSv2, HMDB51
2022 [272]	ViT [146], Swin Transformer [415]	Recognition tasks	Full FT, Mlp-3, Linear Probing, Partial-1, Sidetune, Bias, Adapter, VPT-shallow, VPT-deep	VTAB-1k
2022 [504]	ViT-B/16 [146]	Video action recognition	Full FT, PartialFine-tuning, TemporalFine-tuning, PromptTuning, AttentionalPooling, LinearProbing, Adapter, ST-Adapter	Kinetics-400, Something-Something-v2 (SSv2), Epic-Kitchens-100 (EK100)
2022 [698]	ResNet [219], ConvNeXt [416], Vision Transformer and Swin Transformer [415]	Image classification, part segmentation tasks	P2P Prompting	ModelNet40, ScanObjectNN, ShapeNetPart
2020 [785]	ResNet [219]	Incremental learning, reinforcement learning, computer vision, imitation learning, NLP question answering, single-task transfer learning	Scratch, FT, Elastic Weight Consolidation (EWC), Parameter Superposition (PSP), Progressive Neural Network (PNN), Piggyback (PB), Residual Adapters (RA), Side-tuning	iCIFAR and iTaskonomy, SQuAD v2, Taskonomy

Table 14: Parameter-Efficient Fine-Tuning (PEFT) Methods in Vision Models (2022-2020)