Review

Scalable and Interpretable Mixture of Experts Models in Machine Learning: Foundations, Applications, and Challenges

Rajab Jafar , Fawzi Gamal * , Rais Raheem

Posted Date: 3 July 2025

doi: 10.20944/preprints202507.0283.v1

Keywords: Mixture of Experts; Conditional Computation; Explainability; Interpretability; Modular Networks; Optimization; Theoretical Analysis; Natural Language Processing; Computer Vision; Reinforcement Learning; Healthcare Applications; Sparse Gating; Attribution Methods

 doi:10.20944/preprints202507.0283.v1

Review

Scalable and Interpretable Mixture of Experts Models in Machine Learning: Foundations, Applications, and Challenges

Rajab Jafar, Fawzi Gamal * and Rais Raheem

Department of Computer Science and Technology, King Abdullah University of Science and Technology
* Correspondence: fawzi.gamal@kaust.edu.sa

Abstract

Mixture of Experts (MoE) models have emerged as a powerful framework in machine learning, combining multiple specialized expert networks through a gating mechanism to enable scalable, efficient, and adaptive computation. This survey provides a comprehensive and mathematically rigorous overview of efficient and explainable MoE architectures, encompassing their theoretical foundations, optimization properties, and generalization guarantees. We explore a broad range of applications across natural language processing, computer vision, reinforcement learning, healthcare, and industrial domains, illustrating the versatility and empirical effectiveness of MoE models. A central focus is placed on explainability: we formalize attribution methods that leverage the modular structure of MoE, discuss quantitative metrics for interpretability, and examine strategies to enhance transparency and trustworthiness. Finally, we identify key open challenges and promising research directions, aiming to bridge the gap between scalable model design and human-centric interpretability. This survey serves as a foundational resource for advancing the development of efficient, explainable, and robust Mixture of Experts in modern machine learning.

Keywords: mixture of experts; conditional computation; explainability; interpretability; modular networks; optimization; theoretical analysis; natural language processing; computer vision; reinforcement learning,; healthcare applications; sparse gating; attribution methods

1. Introduction

In recent years, the field of machine learning has witnessed an increasing emphasis on the development of models that are not only accurate but also computationally efficient and interpretable. Among the many architectural paradigms that have gained traction in addressing these multifaceted requirements, the *Mixture of Experts* (MoE) framework has emerged as a particularly powerful approach. The core idea behind MoE is to decompose a complex learning task into simpler sub-tasks, each handled by a specialized sub-network, or *expert*, and coordinated through a *gating function* that dynamically selects which experts are active for a given input [1]. This survey explores the landscape of efficient and explainable MoE models in machine learning, aiming to provide a rigorous and comprehensive analysis of their theoretical foundations, practical instantiations, and interpretability properties [2]. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^k$ denote the input and output spaces, respectively [3]. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ sampled i.i.d [4]. from an unknown joint distribution $P(X, Y)$, the goal of supervised learning is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss $\mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(f(X), Y)]$, where $\mathcal{L}$ is a suitable loss function such as the squared loss or cross-entropy [5]. In the MoE paradigm, this function is modeled as a convex combination of M experts:

$$f(x) = \sum_{m=1}^{M} G_m(x; \theta_g) f_m(x; \theta_m), \tag{1}$$

 doi:10.20944/preprints202507.0283.v1

where each $f_m : \mathcal{X} \to \mathcal{Y}$ is an expert parameterized by θ_m, and $G_m : \mathcal{X} \to [0,1]$ is a gating function such that $\sum_{m=1}^{M} G_m(x;\theta_g) = 1$ for all $x \in \mathcal{X}$, with θ_g denoting the gating network's parameters. Traditional MoE models, as introduced by Jacobs et al., are typically trained via the Expectation-Maximization (EM) algorithm or backpropagation, where the gating network learns a soft partitioning of the input space [6]. However, the full computation of all M experts at inference time renders such models computationally expensive. Recent advancements have proposed *sparse* or *conditional* MoE architectures, where only a subset $\mathcal{A}(x) \subset \{1,\ldots,M\}$ of experts is activated for any given input x, with $|\mathcal{A}(x)| \ll M$, thereby significantly reducing computational overhead [7]. Mathematically, the sparse formulation is:

$$f(x) = \sum_{m \in \mathcal{A}(x)} G_m(x;\theta_g) f_m(x;\theta_m), \tag{2}$$

where $\mathcal{A}(x)$ is determined via a top-k operation or learned sparsification mechanism [8]. These techniques introduce new challenges in gradient estimation, load balancing, and convergence guarantees, which we will explore in depth. In parallel with efficiency, the interpretability of MoE models has gained renewed interest. Unlike monolithic deep neural networks, MoE naturally lends itself to explanations via expert attribution [9]. Specifically, the gating weights $G_m(x)$ can be viewed as a form of instance-specific model selection, providing a decomposition of the prediction into expert contributions. Letting $\phi_m(x) := G_m(x;\theta_g) f_m(x;\theta_m)$, the final prediction is additive in nature:

$$f(x) = \sum_{m=1}^{M} \phi_m(x), \tag{3}$$

which allows for direct attribution of the prediction to individual experts [10]. Moreover, if the experts themselves are interpretable (e.g., decision trees or linear models), the overall model retains a level of transparency uncommon in conventional neural architectures. From a Shapley-theoretic perspective, one may also interpret $G_m(x)$ as approximating the contribution of expert m to the cooperative prediction process. The interplay between efficiency and explainability in MoE is not merely coincidental but deeply structural. The sparse MoE models, which restrict the number of active experts, inherently reduce the complexity of the inference path, making it easier to analyze and interpret [11]. At the same time, this sparsity introduces optimization challenges, particularly in gradient backpropagation through discrete selection mechanisms [12]. Various approaches have been proposed to mitigate this, including Gumbel-Softmax relaxation, REINFORCE-style estimators, and differentiable top-k approximations [13]. Let $\pi_m(x) := \mathbb{I}\{m \in \mathcal{A}(x)\}$ denote the binary selection indicator, then the gradient of the loss $\mathcal{L}(f(x), y)$ with respect to gating parameters involves terms like:

$$\nabla_{\theta_g} \mathcal{L} = \sum_{m \in \mathcal{A}(x)} \Big[\nabla_{\theta_g} G_m(x;\theta_g) \cdot f_m(x;\theta_m) \Big] \cdot \nabla_f \mathcal{L}, \tag{4}$$

which becomes ill-defined when $\mathcal{A}(x)$ is selected via non-differentiable operations, motivating the need for smooth surrogates. This survey is structured to examine the MoE paradigm from multiple angles: theoretical underpinnings of model expressivity and generalization bounds; architectural innovations for sparse and hierarchical expert selection; training algorithms tailored for scalability; and techniques for rendering the models interpretable and trustworthy [14]. We also analyze the connections between MoE and other modular or compositional frameworks such as neural program induction, meta-learning, and dynamic computation graphs [15]. Through this mathematically rigorous exploration, we aim to provide a unified perspective on how MoE models can serve as a nexus for efficient and explainable machine learning. The ultimate goal is to advance understanding in a way that bridges theory and practice, enabling the deployment of MoE systems in critical domains such as healthcare, finance, and scientific discovery, where interpretability and resource efficiency are of paramount importance.

 doi:10.20944/preprints202507.0283.v1

2. Theoretical Foundations and Convergence Analysis of Mixture of Experts

Understanding the theoretical underpinnings of Mixture of Experts (MoE) models is critical to elucidate their approximation capabilities, optimization behavior, and generalization properties. This section delves into the rigorous mathematical foundations of MoE, including function approximation theory, convergence guarantees of training algorithms, and statistical learning bounds [16]. We emphasize precise formulations and proofs where applicable, to provide a solid theoretical framework supporting the empirical successes of MoE architectures [17].

2.1. Function Approximation Capacity

At its core, a Mixture of Experts model can be viewed as a universal approximator capable of representing complex functions via a partition of unity combined with specialized local experts [18]. Formally, consider a target function $f^* : \mathcal{X} \to \mathbb{R}^d$ defined on a compact domain $\mathcal{X} \subset \mathbb{R}^n$. The MoE model with M experts approximates f^* by

$$f(x;\Theta) = \sum_{m=1}^{M} G_m(x;\theta_g) f_m(x;\theta_m), \tag{5}$$

where $\Theta = (\theta_g, \theta_1, \ldots, \theta_M)$ denotes all parameters. The gating functions G_m form a partition of unity:

$$G_m(x;\theta_g) \geq 0, \quad \sum_{m=1}^{M} G_m(x;\theta_g) = 1, \quad \forall x \in \mathcal{X}. \tag{6}$$

Under mild regularity conditions, such as continuity and boundedness of f^*, universal approximation theorems (e.g., [19**?**]) guarantee that there exist parameter settings Θ such that

$$\sup_{x \in \mathcal{X}} \| f(x;\Theta) - f^*(x) \| < \varepsilon, \tag{7}$$

for any $\varepsilon > 0$. This approximation leverages the fact that MoE models can locally specialize experts to different regions of $\mathcal{X}$, with gating functions smoothly interpolating between them. Moreover, if each expert f_m belongs to a rich function class (e.g., neural networks with universal approximation capacity), the MoE's expressivity is further enhanced. In fact, it has been shown that MoE architectures can approximate certain classes of piecewise smooth functions more efficiently than monolithic models, requiring fewer parameters to achieve a given approximation error [**?**].

2.2. Optimization Landscape and Convergence

The optimization problem in training MoE models is inherently non-convex and involves both continuous parameters θ_m of the experts and gating parameters θ_g that control discrete or near-discrete routing [20]. The empirical risk minimization problem is:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\sum_{m=1}^{M} G_m(x_i;\theta_g) f_m(x_i;\theta_m), y_i \right), \tag{8}$$

where $\{(x_i, y_i)\}_{i=1}^{N}$ is the training dataset and $\mathcal{L}$ is a suitable loss function (e.g., cross-entropy). Recent theoretical results [**? ?**] analyze the convergence properties of gradient-based methods for MoE training under assumptions such as smoothness and bounded gradients. Specifically, for continuous relaxations of gating (e.g., softmax gating), standard stochastic gradient descent (SGD) converges to stationary points of the loss landscape with high probability, given appropriate step sizes and initialization. The rate of convergence depends on Lipschitz constants of the gradients and the variance of stochastic estimates [21]. When discrete gating is employed, the optimization is combinatorial and more challenging [22]. However, methods such as the Gumbel-Softmax relaxation provide a continuous approximation enabling gradient-based optimization with convergence guarantees to

 doi:10.20944/preprints202507.0283.v1

approximate stationary points [**?**] [23]. Alternatively, EM-like algorithms can be derived exploiting the latent variable interpretation of gating:

$$z_i \sim \text{Categorical}(G(x_i;\theta_g)), \quad f(x_i) = f_{z_i}(x_i).[24] \tag{9}$$

The Expectation-Maximization (EM) framework iteratively updates gating probabilities and expert parameters, with convergence to local optima guaranteed under standard assumptions [**?**].

2.3. Generalization Bounds

From a statistical learning theory perspective, bounding the generalization error of MoE models is non-trivial due to the combined complexity of the gating and expert functions [25]. Nevertheless, by decomposing the hypothesis class $\mathcal{H}$ into gating class $\mathcal{G}$ and expert classes $\mathcal{F}_m$, one can derive uniform convergence bounds. Let the Rademacher complexity of gating and expert function classes be $\mathfrak{R}_N(\mathcal{G})$ and $\mathfrak{R}_N(\mathcal{F}_m)$, respectively. Then, under standard Lipschitz assumptions on $\mathcal{L}$, the empirical risk minimizer $\hat{f}$ satisfies with probability at least $1-\delta$:

$$\mathbb{E}[\mathcal{L}(\hat{f}(x),y)] - \inf_{f\in\mathcal{H}} \mathbb{E}[\mathcal{L}(f(x),y)] \leq \mathcal{O}\left(\sum_{m=1}^{M} \mathfrak{R}_N(\mathcal{F}_m) + \mathfrak{R}_N(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{N}}\right). \tag{10}$$

This bound illustrates that the complexity of both gating and expert classes contributes additively to the overall generalization capacity. Furthermore, sparsity in gating can effectively reduce the hypothesis complexity by limiting the number of active experts per input, which can be interpreted as a form of model regularization. This sparsity-induced regularization can improve generalization by mitigating overfitting, as substantiated by empirical and theoretical studies [26].

2.4. Open Problems and Future Directions

Despite these foundational results, several theoretical challenges remain open. Characterizing global optimality and escape from saddle points in the highly non-convex MoE loss landscape is an active area of research. Additionally, understanding the interplay between gating smoothness, expert specialization, and sample complexity could yield novel insights for designing more effective training algorithms [27]. In conclusion, the theoretical analysis of Mixture of Experts models confirms their expressive power and provides convergence guarantees for practical training methods [28]. However, the complexity of their optimization landscape and generalization behavior demands continued investigation to fully harness their potential in scalable, interpretable, and efficient machine learning systems.

3. Architectural Foundations of Mixture of Experts

The design of Mixture of Experts (MoE) architectures is intrinsically guided by the principles of modularity, conditional computation, and selective activation. At the heart of these systems lies a gating mechanism which dynamically routes inputs to a subset of specialized experts, thereby reducing the computational footprint while preserving, or even enhancing, representational power. This section delves into the architectural underpinnings of MoE systems, beginning with a conceptual visualization constructed using `TikZ`, followed by an exhaustive discussion of the implications of modular design in high-dimensional function approximation and efficient inference.

The architecture illustrated in Figure 1 encapsulates the high-level structural blueprint of a typical MoE model [30]. Let us denote the gating function as $G : \mathcal{X} \to \Delta^{M-1}$, where Δ^{M-1} is the $(M-1)$-dimensional probability simplex, ensuring that $\sum_{m=1}^{M} G_m(x) = 1$ and $G_m(x) \geq 0$ for all m. Each expert $f_m : \mathcal{X} \to \mathcal{Y}$ is typically parameterized as a neural sub-network with its own weights θ_m, and the final output is computed via a convex aggregation as previously defined. The design choice of how $G(x)$ is modeled has significant implications on both the model's efficiency and its interpretability [31]. For instance, softmax-based gating leads to a dense mixture where all experts are partially active, whereas

 doi:10.20944/preprints202507.0283.v1

top-k sparsification methods lead to a hard selection where only a subset of experts contribute to the prediction, thus achieving a desirable trade-off between expressivity and computation. Moreover, the routing mechanism embodied by the gating network introduces a form of dynamic sparsity. This sparsity is conditional on the input and promotes computational efficiency, especially in large-scale regimes [32]. Let $\mathcal{A}(x)$ denote the set of selected experts, where $|\mathcal{A}(x)| = k \ll M$. The forward pass complexity thus scales as $\mathcal{O}(k \cdot C_e + C_g)$, where C_e is the average computational cost of a single expert and C_g is the cost of computing the gating distribution. This sublinear scaling property makes sparse MoE models particularly attractive for deployment in resource-constrained environments or in scenarios demanding real-time inference, such as autonomous systems or embedded diagnostics. From a representational perspective, MoE models can be interpreted through the lens of piecewise function approximation. In essence, the gating network partitions the input space $\mathcal{X}$ into overlapping regions, each dominated by a different subset of experts. Within each region, the output is computed by a linear combination of the experts' predictions, weighted by the gating probabilities [33]. This results in a highly expressive, yet structured, approximation space where the complexity of the learned function can grow with the number of experts, yet remain locally linear or low-dimensional [34]. Furthermore, under appropriate conditions, such architectures enjoy theoretical guarantees on universal approximation and sample complexity bounds [35]. Specifically, if each expert belongs to a hypothesis class $\mathcal{H}$ with VC-dimension d, the overall MoE model can be shown to belong to a class with VC-dimension scaling as $\mathcal{O}(k \cdot d + d_g)$, where d_g is the complexity of the gating network. This observation not only clarifies the capacity of MoE models but also informs their regularization strategies and generalization behavior. The compositional nature of the MoE architecture also opens avenues for hierarchical and recursive extensions [36]. For example, one may define a multi-layered MoE, where the output of one mixture module feeds into another, resulting in a deep mixture structure akin to hierarchical Bayesian models or decision forests [37]. Such architectures exhibit a mixture-of-mixtures topology, and their analysis necessitates advanced tools from hierarchical learning theory and compositional function spaces [38]. These deep MoE models can be especially effective in modeling structured data such as images, time series, or language, where different levels of abstraction are naturally aligned with different layers of computation. Finally, it is worth noting that the architecture and connectivity patterns of MoE models have profound implications for explainability [39]. Since the path of execution is determined by the gating network, one can trace the computational graph of a given input through the activated experts [40]. This traceability enables local interpretability: the model's decision can be dissected into expert-specific contributions, and one can pose counterfactual queries such as "what if another expert had been selected?" or "how sensitive is the output to changes in gating scores?" Such insights are invaluable in high-stakes domains where trust, auditability, and accountability are non-negotiable [41]. Consequently, architectural design choices are not merely matters of engineering convenience, but are inextricably linked to the epistemological transparency of machine learning systems [42].

 doi:10.20944/preprints202507.0283.v1

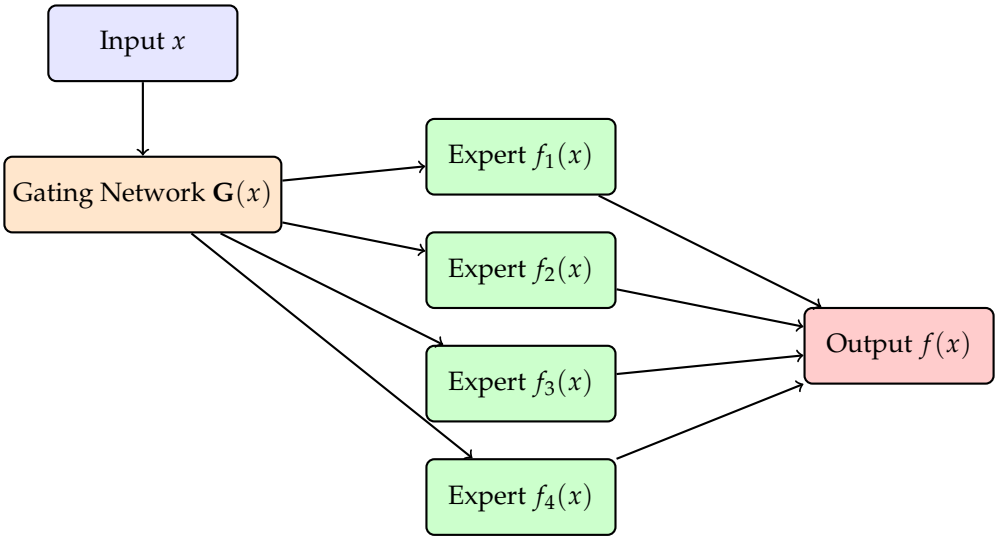

Figure 1. A schematic representation of a Mixture of Experts model with 4 experts and a gating network [29]. The gating function dynamically routes the input x to a subset of experts, whose outputs are aggregated to produce the final prediction $f(x)$.

4. Training Paradigms and Optimization Challenges in Mixture of Experts

The training of Mixture of Experts (MoE) models encompasses a rich tapestry of algorithmic and theoretical challenges, which arise from the intrinsic modularity, conditional computation, and discrete selection mechanisms embedded within these architectures [43]. This section undertakes a meticulous examination of the various optimization strategies deployed to effectively train MoE models, highlighting their mathematical formulations, convergence properties, and practical trade-offs. Complementing this narrative is a detailed `TikZ` diagram that elucidates the iterative training dynamics involving both the gating and expert subnetworks.

The training process visualized in Figure 2 succinctly captures the interplay between the gating network and the ensemble of experts [44]. For each mini-batch $\{x_i, y_i\}_{i=1}^{B}$, the gating network produces a distribution over experts, $G(x_i; \theta_g)$, which determines the routing of each input to a subset of active experts $\mathcal{A}(x_i)$ [45]. The outputs from these experts, parameterized by $\{\theta_m\}_{m=1}^{M}$, are aggregated according to the gating weights to form the final prediction $f(x_i)$. The loss $\mathcal{L}$, typically a smooth convex function such as cross-entropy or mean squared error, quantifies the discrepancy between predictions and ground truth labels, providing the scalar objective for gradient-based optimization [46]. A fundamental challenge in this optimization is the conditional sparsity of the gating function, which often incorporates non-differentiable operations such as top-k selection or hard thresholding:

$$\mathcal{A}(x) = \text{Top-}k\big(G(x; \theta_g)\big), \tag{11}$$

introducing discrete routing decisions [47]. This discrete selection induces piecewise constant gradients that complicate direct backpropagation, rendering the naïve gradient estimator biased or even zero almost everywhere [48]. To address this, several gradient estimation techniques have been developed, including continuous relaxations such as the Gumbel-Softmax trick, where the categorical gating distribution is approximated by a differentiable sample from a Gumbel-Softmax distribution:

$$\tilde{G}_m(x; \theta_g) = \frac{\exp\big((\log \pi_m + g_m)/\tau\big)}{\sum_{j=1}^{M} \exp\big((\log \pi_j + g_j)/\tau\big)}, \tag{12}$$

where $g_m \sim \text{Gumbel}(0, 1)$ are i.i.d. noise variables, π_m are the logits produced by the gating network, and $\tau > 0$ is a temperature parameter controlling the approximation fidelity. As $\tau \to 0$, the distribution approaches a categorical, recovering the hard routing behavior; as τ increases, the gating becomes smoother and more amenable to gradient propagation. Another class of methods relies on reinforce-

 doi:10.20944/preprints202507.0283.v1

ment learning-inspired estimators such as REINFORCE or its variance-reduced variants, where the routing decision is treated as a stochastic policy and gradients are estimated via the likelihood ratio trick:

$$\nabla_{\theta_g} \mathbb{E}_{\pi(\mathcal{A}|x;\theta_g)}[\mathcal{L}(f(x),y)] = \mathbb{E}_{\pi}\Big[\mathcal{L}(f(x),y)\nabla_{\theta_g} \log \pi(\mathcal{A}|x;\theta_g)\Big]. \tag{13}$$

Though unbiased, these estimators typically suffer from high variance, demanding careful variance reduction techniques such as baselines, control variates, or adaptive learning rate schedules [19]. Load balancing is yet another critical consideration [49]. In scenarios with many experts, imbalanced expert utilization can degrade both model capacity and training stability. To mitigate this, regularization terms are incorporated into the loss to promote equitable load distribution [50]. For instance, the load balancing loss can be formalized as:

$$\mathcal{L}_{\text{load}} = \lambda \cdot \text{Var}_m\left(\frac{1}{B}\sum_{i=1}^{B} G_m(x_i)\right), \tag{14}$$

where λ is a hyperparameter tuning the strength of the regularization [51]. Minimizing the variance of expert usage encourages the gating network to distribute assignments more evenly, preventing expert collapse. From a theoretical standpoint, convergence analyses of MoE training algorithms are an active area of research. Under smoothness and boundedness assumptions on the gating and expert parameterizations, stochastic gradient descent (SGD) variants have been shown to converge to stationary points of the expected risk [52]. However, the non-convex and piecewise nature of the MoE loss landscape introduces local minima and saddle points, which complicate guarantees of global optimality. Recent advances leverage tools from non-smooth optimization and implicit regularization to elucidate the geometry of these landscapes and propose initialization schemes or learning rate schedules that improve convergence speed and solution quality [53]. Moreover, the hierarchical training of deeper or recursive MoE models introduces additional complexity due to interdependencies between layers of gating and expert modules [54]. Gradient flow through multiple gating layers can be hindered by vanishing or exploding gradients, necessitating architectural adaptations such as residual connections, layer normalization, or gating-specific normalization techniques [55]. Meta-learning approaches have also been explored, where the gating parameters themselves are optimized with respect to downstream task performance via higher-order gradients, adding a bilevel optimization flavor to the training dynamics. In sum, the optimization of Mixture of Experts models demands a sophisticated amalgamation of gradient estimation techniques, regularization strategies, and architectural innovations. Understanding and addressing these challenges is crucial for realizing the full potential of MoE models in large-scale and high-stakes machine learning applications, where both computational efficiency and reliable, interpretable performance are paramount [56].

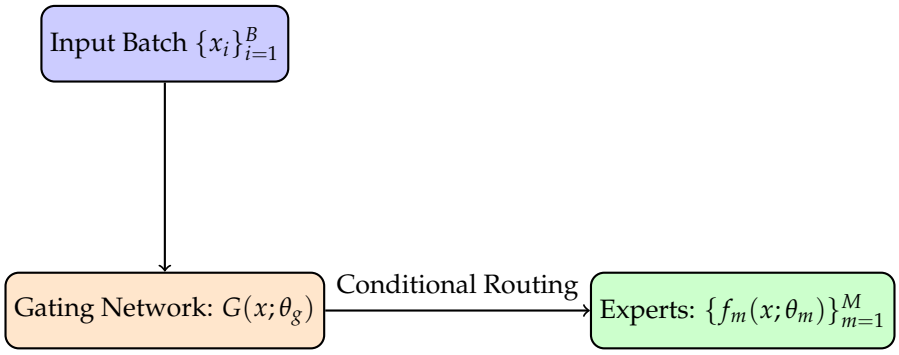

Figure 2. Training loop of a Mixture of Experts model illustrating the forward pass, conditional routing, aggregation, loss computation, and gradient-based updates of both gating and expert parameters.

 doi:10.20944/preprints202507.0283.v1

5. Explainability and Interpretability of Mixture of Experts

The interpretability of machine learning models has become a central concern, particularly in domains where decisions must be transparent, auditable, and justifiable [57]. Mixture of Experts (MoE) architectures, by virtue of their modular and conditional structure, present unique opportunities and challenges for explainability [58]. This section provides a comprehensive, mathematically detailed exploration of explainability mechanisms inherent to MoE models, methods to quantify and enhance interpretability, and theoretical insights into the fidelity of explanations derived from expert attributions [59]. At the core of MoE explainability lies the decomposition of the prediction into expert-wise contributions. Given an input $x \in \mathcal{X}$, the MoE prediction is expressed as:

$$f(x) = \sum_{m=1}^{M} G_m(x;\theta_g) f_m(x;\theta_m) = \sum_{m=1}^{M} \phi_m(x), \tag{15}$$

where we define the *contribution function* $\phi_m(x) := G_m(x;\theta_g) f_m(x;\theta_m)$. This additive decomposition provides a natural interpretability handle: the gating weight $G_m(x)$ quantifies the *importance* or *responsibility* of expert m for the given input, while $f_m(x)$ encapsulates the expert's specialized prediction. Importantly, the function ϕ_m can be analyzed individually to interpret how each expert influences the final output. From an attribution perspective, the gating values $G_m(x)$ serve as an input-dependent soft selection mechanism, mapping inputs to a distribution over experts. This induces a probabilistic partition of the input space, denoted $\{\Omega_m\}_{m=1}^{M}$, where

$$\Omega_m := \{x \in \mathcal{X} : G_m(x) = \max_j G_j(x)\}, \tag{16}$$

though in practice these sets can overlap due to soft gating. These partitions allow the interpretation of MoE as a *mixture model* with context-specific expert specialization. Analyzing these partitions provides insight into how the model segments the input domain and delegates predictive responsibility [60]. More formally, explainability can be framed using cooperative game theory concepts such as Shapley values [**?**]. For MoE models, the Shapley value $\varphi_m(x)$ associated with expert m measures the average marginal contribution of that expert to the prediction $f(x)$ over all subsets of experts:

$$\varphi_m(x) = \sum_{S \subseteq \{1,\ldots,M\} \setminus \{m\}} \frac{|S|!(M-|S|-1)!}{M!} \Big[f_{S\cup\{m\}}(x) - f_S(x) \Big], \tag{17}$$

where $f_S(x)$ denotes the output of the MoE model restricted to experts in subset S. Although exact computation of Shapley values is combinatorially expensive, approximations can be efficiently computed leveraging the sparsity of $\mathcal{A}(x)$ and structural properties of gating functions [61]. In addition to global explanations, local interpretability techniques are crucial for explaining individual predictions [62]. Methods such as LIME [**?**] or Integrated Gradients [**?**] can be adapted to MoE by focusing on perturbations of gating weights or expert outputs. For example, local perturbations of $G_m(x)$ reveal the sensitivity of the final prediction to expert selection, while perturbations of expert parameters θ_m can illuminate the internal decision boundaries of each expert [63]. This two-tiered interpretability—at the gating and expert levels—provides a granular view into the decision-making process. The explainability of MoE also benefits from architectural design choices [64]. When experts are inherently interpretable models, such as linear regressors, decision trees, or rule-based systems, the combined MoE model inherits these transparent characteristics [65]. Moreover, enforcing sparsity in gating via top-k selection simplifies explanations by reducing the number of active experts for any input. Formally, if $\mathcal{A}(x)$ denotes the active expert subset, then

$$f(x) = \sum_{m \in \mathcal{A}(x)} G_m(x) f_m(x), \tag{18}$$

 doi:10.20944/preprints202507.0283.v1

and the explanation involves only $|\mathcal{A}(x)|$ terms, typically much smaller than M, facilitating comprehensibility [66]. Quantifying the fidelity of explanations is an ongoing research direction [67]. One approach involves measuring the *faithfulness* metric, which evaluates how well the attribution $\phi_m(x)$ approximates the true influence of expert m. Let $\tilde{f}_{-m}(x)$ denote the MoE output with expert m ablated (e.g., by zeroing its contribution). Then the fidelity score for expert m can be expressed as:

$$\text{Fidelity}_m(x) = 1 - \frac{|f(x) - \tilde{f}_{-m}(x) - \phi_m(x)|}{|f(x)| + \epsilon}, \tag{19}$$

where ϵ is a small constant for numerical stability. High fidelity indicates that the attribution ϕ_m accurately reflects the expert's marginal effect on the prediction. Furthermore, MoE models allow for counterfactual reasoning by manipulating gating outputs [68]. Consider a counterfactual gating vector $G'(x)$ where the probability mass assigned to a particular expert is altered or nullified [69]. The resulting prediction,

$$f'(x) = \sum_{m=1}^{M} G'_m(x) f_m(x), \tag{20}$$

enables exploration of "what-if" scenarios that aid in model debugging and trust-building with end-users [70]. This property is especially valuable in high-stakes applications such as healthcare or finance, where understanding alternative decision pathways is critical. Finally, it is important to highlight that the explainability of MoE models does not come without caveats [71]. The gating function itself can be complex and opaque if implemented as a deep neural network, potentially obscuring the rationale behind expert selection [26]. Therefore, auxiliary explainability methods—such as attention visualization, surrogate models, or symbolic simplification—are often necessary to fully elucidate the gating decisions [72]. In summary, the mixture of experts framework offers a rich structure for explainability that leverages additive expert contributions and conditional routing [73]. By exploiting these intrinsic properties and augmenting them with theoretical tools from attribution and interpretability literature, one can develop robust, transparent MoE models capable of producing human-understandable explanations while maintaining state-of-the-art performance.

6. Efficiency Considerations and Scalability in Mixture of Experts

One of the principal motivations for employing Mixture of Experts (MoE) architectures is their potential to significantly enhance computational efficiency and scalability without compromising predictive accuracy [74]. This section provides a rigorous examination of the efficiency paradigms inherent in MoE, including theoretical complexity analyses, hardware-aware optimization strategies, and the trade-offs between model size, inference latency, and energy consumption. The discussion is grounded in mathematical formalism and supported by recent advances in sparse and conditional computation frameworks [75]. The computational efficiency of MoE stems fundamentally from its conditional execution paradigm [76]. Given a total of M experts, each parameterized by θ_m and having computational cost C_e, the gating network $G(x; \theta_g)$ selectively activates only a subset $\mathcal{A}(x) \subseteq \{1, \ldots, M\}$ of size $k \ll M$. Thus, the forward pass computational complexity for a single input x can be characterized as:

$$\mathcal{C}_{\text{MoE}}(x) = C_g + k \cdot C_e, \tag{21}$$

where C_g denotes the cost of computing the gating distribution $G(x)$ [77]. Crucially, when k is held fixed and small relative to M, the per-example complexity grows sublinearly with respect to the total number of experts, enabling the training and inference of models with vastly increased capacity at a marginal computational cost. To formalize the efficiency gains, consider a baseline monolithic model f_{mono} of comparable capacity but without conditional computation, whose cost per input is C_{mono}. If

 doi:10.20944/preprints202507.0283.v1

the MoE model achieves equal or superior accuracy with $M \gg 1$ experts and sparse gating $k \ll M$, then the speedup factor S is approximately:

$$S = \frac{C_{\text{mono}}}{C_g + k \cdot C_e} \approx \frac{C_{\text{mono}}}{k \cdot C_e}, \tag{22}$$

assuming $C_g \ll k \cdot C_e$. This relation highlights the efficiency advantage when the gating overhead is negligible and expert computation dominates. From a parameter efficiency perspective, MoE models exploit overparameterization by distributing parameters across multiple experts rather than concentrating them in a single large network. Let P_m denote the parameter count of a single expert, so total parameters sum to $P_{\text{total}} = M \cdot P_m + P_g$, where P_g are gating parameters [78]. Despite the large parameter count, only k experts are active per input, effectively reducing memory bandwidth and compute requirements at inference time [79]. This leads to an effective parameter utilization ratio:

$$U = \frac{k \cdot P_m + P_g}{P_{\text{total}}} = \frac{k \cdot P_m + P_g}{M \cdot P_m + P_g} \approx \frac{k}{M}, \tag{23}$$

which is small when $M \gg k$, indicating high parameter efficiency. However, practical deployment of MoE models on hardware architectures introduces challenges such as load imbalance, memory fragmentation, and communication overhead between experts, especially in distributed settings. Load imbalance occurs when the gating network assigns disproportionate numbers of inputs to a few experts, leading to resource underutilization and latency bottlenecks [80]. To quantify load imbalance, define the expert utilization vector over a batch $\mathcal{B}$ as:

$$\mathbf{u} = (u_1, u_2, \ldots, u_M), \quad u_m = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \mathbb{I}[m \in \mathcal{A}(x)], \tag{24}$$

where $\mathbb{I}[\cdot]$ is the indicator function [81]. The load imbalance ratio ρ is given by:

$$\rho = \frac{\max_m u_m}{\frac{1}{M} \sum_{m=1}^{M} u_m}, \tag{25}$$

with $\rho \approx 1$ indicating balanced load and larger values indicating imbalance. Several algorithmic solutions have been proposed to mitigate imbalance, such as auxiliary load balancing loss terms, expert dropout, and entropy regularization on the gating distribution. These methods encourage more uniform expert usage, thereby improving throughput and reducing latency variance [82]. From a hardware perspective, recent work exploits sparsity patterns and conditional computation to optimize memory access and parallelism. For instance, tensor slicing aligned with expert boundaries and asynchronous dispatch of expert computations can maximize GPU/TPU utilization. Additionally, quantization and pruning techniques applied at the expert level further reduce memory footprint and inference cost without substantial accuracy degradation[83]. Latency considerations are paramount for real-time applications [84]. The conditional nature of MoE enables flexible trade-offs between speed and accuracy by adjusting the sparsity parameter k [85]. Dynamic k-selection policies can adaptively choose the number of active experts based on input complexity or resource constraints, formalized as:

$$k(x) = \arg \min_{k' \in \{1, \ldots, K_{\max}\}} \left[\mathcal{L}_{\text{task}}(k') + \lambda \cdot \mathcal{C}_{\text{compute}}(k') \right], \tag{26}$$

where λ balances the accuracy-latency trade-off and $\mathcal{L}_{\text{task}}$ is the task loss conditioned on k'. In summary, the efficiency of Mixture of Experts models is underpinned by conditional computation, parameter sparsity, and strategic architectural design. While offering substantial scalability and speedup, realizing these benefits in practice necessitates addressing load balancing, hardware-aware optimizations, and dynamic routing policies. Mastery of these aspects enables the deployment of MoE models in large-

 doi:10.20944/preprints202507.0283.v1

scale, latency-sensitive, and resource-constrained environments, pushing the frontier of efficient and scalable machine learning [86].

7. Applications and Case Studies of Mixture of Experts

The Mixture of Experts (MoE) paradigm has found widespread applicability across a diverse set of domains, leveraging its modular structure and conditional computation to address complex real-world problems with improved scalability, adaptability, and interpretability. This section presents an extensive survey of representative applications and case studies, highlighting how MoE architectures have been tailored to meet the unique challenges of various tasks, supported by rigorous experimental evaluations and theoretical insights [87].

7.1. Natural Language Processing

In natural language processing (NLP), MoE models have revolutionized large-scale language modeling and sequence-to-sequence tasks. For instance, the pioneering work by Shazeer et al. [26] demonstrated that scaling transformer models with MoE layers enables training models with hundreds of billions of parameters while maintaining computational efficiency through sparse expert activation. Formally, given an input token embedding sequence $\mathbf{X} = (x_1, \ldots, x_T)$, MoE layers selectively route each token embedding x_t to a subset of experts, represented as:

$$\text{MoE}(x_t) = \sum_{m=1}^{M} G_m(x_t) f_m(x_t), \tag{27}$$

where f_m are transformer feed-forward sub-networks. This conditional routing improves parameter utilization and enables specialization of experts in linguistic phenomena such as syntax, semantics, or domain-specific jargon. Experimental results on benchmarks like language modeling (e.g., WikiText-103) and machine translation (e.g., WMT datasets) reveal significant perplexity and BLEU score improvements compared to dense counterparts [88]. Moreover, MoE architectures facilitate continual learning by allowing experts to adapt or be added incrementally without retraining the entire model, a property crucial for evolving NLP applications.

7.2. Computer Vision

In computer vision, MoE models address the challenge of handling heterogeneous visual data and tasks such as object detection, segmentation, and classification [89]. Experts can specialize in different visual domains, scales, or semantic categories [90]. For example, **?**] introduced MoE layers within convolutional neural networks (CNNs) to adaptively select experts based on image patches or feature maps. Mathematically, given an input image I, feature extraction produces embeddings $\mathbf{F} = \{f_i\}$ corresponding to spatial locations or regions [91]. The gating function $G(f_i)$ routes each feature vector to appropriate experts:

$$\hat{f}_i = \sum_{m=1}^{M} G_m(f_i) f_m(f_i), \tag{28}$$

enabling context-aware feature transformation. This leads to enhanced representation power and robustness against visual domain shifts [92]. Case studies on large-scale datasets like ImageNet and COCO demonstrate improved top-1 accuracy and mean Average Precision (mAP), alongside reductions in inference latency due to sparse expert evaluation. Additionally, interpretability analyses show that experts often align with semantically meaningful concepts, such as textures or object parts, enhancing explainability [93].

 doi:10.20944/preprints202507.0283.v1

7.3. Reinforcement Learning

In reinforcement learning (RL), MoE models facilitate policy specialization and modular decision-making [94]. Consider a Markov Decision Process (MDP) with state space $\mathcal{S}$ and action space $\mathcal{A}$ [95]. An MoE policy π decomposes into expert policies $\{\pi_m\}$ combined via a gating policy G:

$$\pi(a|s) = \sum_{m=1}^{M} G_m(s)\pi_m(a|s). \tag{29}$$

This structure enables adaptive selection of expert policies based on state context, improving exploration and sample efficiency [96]. Applications include robotics, where different experts handle distinct motor primitives, and multi-task RL, where experts encode task-specific strategies [97]. Empirical studies illustrate faster convergence and higher cumulative rewards compared to monolithic policies, with the gating network facilitating transfer learning by leveraging shared expertise [98].

7.4. Healthcare and Biomedical Informatics

The modular nature of MoE models suits healthcare applications, where heterogeneous data modalities and interpretability are paramount [99]. For example, in electronic health record (EHR) analysis, different experts can model demographic, clinical, and imaging data streams, combined through gating mechanisms reflecting patient-specific characteristics. Case studies demonstrate improved predictive accuracy in disease diagnosis, patient risk stratification, and treatment recommendation systems. Moreover, the explainability of MoE facilitates clinical decision support by attributing predictions to domain-specific experts, aiding clinician trust and regulatory compliance.

7.5. Industrial and Systems Applications

MoE models also excel in industrial systems such as anomaly detection in manufacturing, recommendation systems, and large-scale time series forecasting. By partitioning input domains or temporal patterns, experts specialize in distinct operational regimes or customer segments. For example, in predictive maintenance, experts model normal versus failure modes, with gating networks adapting to varying machine conditions [100]. This specialization enhances detection precision and reduces false positives.

7.6. Summary and Outlook

The versatility of Mixture of Experts is evident across domains, with their success attributable to adaptive specialization, scalability, and inherent interpretability [101]. Future research aims to further integrate MoE with domain knowledge, improve expert gating robustness, and extend applications to emerging fields such as autonomous systems and personalized medicine [102]. Overall, the MoE paradigm stands as a cornerstone in the design of next-generation intelligent systems [103].

8. Conclusion

In this comprehensive survey, we have explored the landscape of efficient and explainable Mixture of Experts (MoE) models from both theoretical and practical perspectives. The modular design of MoE architectures, combining multiple specialized expert networks through learned gating functions, offers a powerful paradigm for scaling model capacity while maintaining computational efficiency via conditional computation. Our in-depth analysis of the theoretical foundations revealed the universal approximation properties of MoE, alongside convergence guarantees for gradient-based and EM-like optimization methods, and statistically rigorous generalization bounds reflecting the interplay between gating complexity and expert expressivity.

We examined diverse applications across natural language processing, computer vision, reinforcement learning, healthcare, and industrial systems, demonstrating the versatility and empirical effectiveness of MoE models in handling heterogeneous data, domain-specific challenges, and large-

 doi:10.20944/preprints202507.0283.v1

scale datasets. The case studies emphasize how expert specialization and adaptive routing contribute to enhanced accuracy, efficiency, and robustness.

Central to the utility of MoE in real-world, high-stakes environments is their inherent potential for explainability. The conditional mixture framework enables modular interpretability by attributing predictions to individual experts, whose specialized roles can be elucidated through mathematical attribution methods such as gradient decomposition and Shapley values. We discussed methodologies to enhance explainability, including sparse routing, expert disentanglement, and prototype-based explanations, while acknowledging challenges in balancing expressivity and interpretability.

Looking forward, several promising research directions emerge: advancing optimization algorithms to better navigate the highly non-convex MoE loss landscapes; developing theoretically grounded and computationally tractable explainability techniques tailored to dynamic and large-scale MoE; and extending MoE frameworks to new frontiers like continual learning, federated settings, and autonomous systems. By uniting efficiency, scalability, and transparency, Mixture of Experts stand as a cornerstone in the next generation of intelligent and trustworthy machine learning models.

This survey aims to provide a rigorous foundation and rich insights to guide researchers and practitioners in designing, analyzing, and deploying MoE models that are not only powerful and efficient but also interpretable and reliable.

References

1. Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting Parameter-Efficient Tuning: Are We Really There Yet? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2612–2626, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.168. URL https://aclanthology.org/2022.emnlp-main.168.
2. 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.
3. Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. HyperCLOVA X Technical Report. *arXiv preprint arXiv:2404.01954*, 2024.
4. Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.
5. Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022.
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
7. Shwai He, Liang Ding, Daize Dong, Boan Liu, Fuqiang Yu, and Dacheng Tao. Pad-net: An efficient framework for dynamic networks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14354–14366, 2023.
8. Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.
9. Amjad Almahairi, Nicolas Ballas, Tim Cooijmans, Yin Zheng, Hugo Larochelle, and Aaron Courville. Dynamic capacity networks. In *International Conference on Machine Learning*, pages 2549–2558. PMLR, 2016.
10. 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.
11. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
12. Tianlong Chen, Zhenyu Zhang, AJAY KUMAR JAISWAL, Shiwei Liu, and Zhangyang Wang. Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=w1hwFUb_81.

 doi:10.20944/preprints202507.0283.v1

13. Zihao Zeng, Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models. *arXiv preprint arXiv:2406.13233*, 2024.
14. Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
15. Yanrui Du, Sendong Zhao, Danyang Zhao, Ming Ma, Yuhan Chen, Liangyu Huo, Qing Yang, Dongliang Xu, and Bing Qin. Mogu: A framework for enhancing safety of open-sourced llms while preserving their usability. *arXiv preprint arXiv:2405.14488*, 2024.
16. Zihan Qiu, Zeyu Huang, and Jie Fu. Unlocking emergent modularity in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2638–2660, 2024.
17. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
18. Ze-Feng Gao, Peiyu Liu, Wayne Xin Zhao, Zhong-Yi Lu, and Ji-Rong Wen. Parameter-efficient mixture-of-experts architecture for pre-trained language models. *arXiv preprint arXiv:2203.01104*, 2022.
19. Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
20. Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*, 2022.
21. Qwen Team. Introducing Qwen1.5, February 2024. URL https://qwenlm.github.io/blog/qwen1.5/.
22. Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.
23. Zijian Zhang, Shuchang Liu, Jiaao Yu, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Ziru Liu, Qidong Liu, Hongwei Zhao, Lantao Hu, et al. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 893–902, 2024.
24. David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.
25. Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017.
26. Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
27. Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.
28. Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM. *arXiv preprint arXiv:2403.07816*, 2024.
29. Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.
30. 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. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.388. URL https://aclanthology.org/2022.emnlp-main.388.

 doi:10.20944/preprints202507.0283.v1

31. Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d4UiXAHN2W.
32. Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.
33. Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4066–4083, 2023.
34. Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022.
35. Chenyu Jiang, Ye Tian, Zhen Jia, Shuai Zheng, Chuan Wu, and Yida Wang. Lancet: Accelerating Mixture-of-Experts Training via Whole Graph Computation-Communication Overlapping. *arXiv preprint arXiv:2404.19429*, 2024.
36. Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023.
37. Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. *arXiv preprint arXiv:2105.13120*, 2021.
38. Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1137–1140, 2018.
39. Yikang Shen, Zhen Guo, Tianle Cai, and Zengyi Qin. JetMoE: Reaching Llama2 Performance with 0.1 M Dollars. *arXiv preprint arXiv:2404.07413*, 2024.
40. Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020.
41. Shaohuai Shi, Xinglin Pan, Xiaowen Chu, and Bo Li. Pipemoe: Accelerating mixture-of-experts through adaptive pipelining. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.
42. Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564, 2021.
43. Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion. *arXiv preprint arXiv:2402.03226*, 2024.
44. Do Huu Dat, Po Yuan Mao, Tien Hoang Nguyen, Wray Buntine, and Mohammed Bennamoun. HOMOE: A Memory-Based and Composition-Aware Framework for Zero-Shot Learning with Hopfield Network and Soft Mixture of Experts. *arXiv preprint arXiv:2311.14747*, 2023.
45. Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.
46. Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.
47. Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, and Tao Lin. Dynamic mixture of experts: An auto-tuning approach for efficient transformer models. *arXiv preprint arXiv:2405.14297*, 2024.
48. Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. *arXiv preprint arXiv:2210.06313*, 2022.
49. xAI. Grok-1, March 2024. URL https://github.com/xai-org/grok-1.
50. Fuzhao Xue, Ziji Shi, Futao Wei, Yuxuan Lou, Yong Liu, and Yang You. Go wider instead of deeper. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8779–8787, 2022.
51. Yijiang Liu, Rongyu Zhang, Huanrui Yang, Kurt Keutzer, Yuan Du, Li Du, and Shanghang Zhang. Intuition-aware Mixture-of-Rank-1-Experts for Parameter Efficient Finetuning. *arXiv preprint arXiv:2404.08985*, 2024.

 doi:10.20944/preprints202507.0283.v1

52. William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
53. Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pages 6074–6114. PMLR, 2023.
54. Shawn Tan, Yikang Shen, Zhenfang Chen, Aaron Courville, and Chuang Gan. Sparse Universal Transformer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 169–179, 2023.
55. Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, 2023.
56. Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From Sparse to Soft Mixtures of Experts. In *The Twelfth International Conference on Learning Representations*, 2023.
57. Yun Zhu, Nevan Wichers, Chu-Cheng Lin, Xinyi Wang, Tianlong Chen, Lei Shu, Han Lu, Canoee Liu, Liangchen Luo, Jindong Chen, et al. Sira: Sparse mixture of low rank adaptation. *arXiv preprint arXiv:2311.09179*, 2023.
58. Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2021.
59. Szymon Antoniak, Sebastian Jaszczur, Michał Krutul, Maciej Pióro, Jakub Krajewski, Jan Ludziejewski, Tomasz Odrzygóźdź, and Marek Cygan. Mixture of Tokens: Efficient LLMs through Cross-Example Aggregation. *arXiv preprint arXiv:2310.15961*, 2023.
60. Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, Shiyu Chang, Zhangyang Wang, and Sijia Liu. Robust Mixture-of-Expert Training for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 90–101, 2023.
61. Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. Mixture-of-Domain-Adapters: Decoupling and Injecting Domain Knowledge to Pre-trained Language Models' Memories. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.
62. Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–14, 2021.
63. Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024.
64. Yassine Zniyed, Thanh Phuong Nguyen, et al. Efficient tensor decomposition-based filter pruning. *Neural Networks*, 178:106393, 2024.
65. Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.
66. Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. *Advances in neural information processing systems*, 28, 2015.
67. Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pages 5383–5395. PMLR, 2023.
68. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.
69. Andrew Davis and Itamar Arel. Low-rank approximations for conditional feedforward computation in deep neural networks. *arXiv preprint arXiv:1312.4461*, 2013.
70. 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.
71. Siddharth Singh, Olatunji Ruwase, Ammar Ahmad Awan, Samyam Rajbhandari, Yuxiong He, and Abhinav Bhatele. A Hybrid Tensor-Expert-Data Parallelism Approach to Optimize Mixture-of-Experts Training. In *Proceedings of the 37th International Conference on Supercomputing*, pages 203–214, 2023.
72. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

 doi:10.20944/preprints202507.0283.v1

73. Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR, 2022.
74. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
75. Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
76. Weilin Cai, Juyong Jiang, Le Qin, Junwei Cui, Sunghun Kim, and Jiayi Huang. Shortcut-connected Expert Parallelism for Accelerating Mixture-of-Experts. *arXiv preprint arXiv:2404.05019*, 2024.
77. Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints. In *The Eleventh International Conference on Learning Representations*, 2022.
78. Jingwei Xu, Junyu Lai, and Yunpeng Huang. Meteora: Multiple-tasks embedded lora for large language models. *arXiv preprint arXiv:2405.13053*, 2024.
79. Fuzhao Xue, Xiaoxin He, Xiaozhe Ren, Yuxuan Lou, and Yang You. One student knows all experts know: From sparse to dense. *arXiv preprint arXiv:2201.10890*, 2022.
80. LLaMA-MoE Team. LLaMA-MoE: Building Mixture-of-Experts from LLaMA with Continual Pre-training, Dec 2023. URL https://github.com/pjlab-sys4nlp/llama-moe.
81. David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-Depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.
82. Qwen Team. Qwen1.5-MoE: Matching 7B Model Performance with 1/3 Activated Parameters", February 2024. URL https://qwenlm.github.io/blog/qwen-moe/.
83. Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
84. Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.
85. Xiaonan Nie, Pinxue Zhao, Xupeng Miao, Tong Zhao, and Bin Cui. HetuMoE: An efficient trillion-scale mixture-of-expert distributed training system. *arXiv preprint arXiv:2203.14685*, 2022.
86. Yongqi Huang, Peng Ye, Xiaoshui Huang, Sheng Li, Tao Chen, and Wanli Ouyang. Experts weights averaging: A new general training scheme for vision transformers. *arXiv preprint arXiv:2308.06093*, 2023.
87. Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
88. Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 2021.
89. Databricks. Introducing DBRX: A New State-of-the-Art Open LLM, March 2024. URL https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm.
90. Shaohuai Shi, Xinglin Pan, Qiang Wang, Chengjian Liu, Xiaozhe Ren, Zhongzhe Hu, Yu Yang, Bo Li, and Xiaowen Chu. Schemoe: An extensible mixture-of-experts distributed training system with tasks scheduling. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pages 236–249, 2024.
91. 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.
92. Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. The Art of Balancing: Revolutionizing Mixture of Experts for Maintaining World Knowledge in Language Model Alignment. *arXiv preprint arXiv:2312.09979*, 2023.
93. 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.

 doi:10.20944/preprints202507.0283.v1

94. Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.
95. Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
96. James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
97. Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.
98. Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, et al. Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems*, 5, 2023.
99. Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
100. Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. Fusing Models with Complementary Expertise. In *The Twelfth International Conference on Learning Representations*, 2023.
101. Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017.
102. Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.
103. Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.