Towards Robust and Scalable Mixture of Experts Architectures for Large Language and Vision Models

Aamina Yousra, Jumanah Fawziya, Fawzi Gamal
Department of Computer Science and Technology,
King Abdullah University of Science and Technology

Abstract

The advent of foundation-scale deep learning models, characterized by unprecedented model sizes and multi-modal capabilities, has revitalized interest in Mixture of Experts (MoE) architectures due to their potential for efficient conditional computation and scalability. However, robustness challenges—including routing instability, expert overload, and vulnerability to distributional shifts and adversarial attacks—pose significant barriers to reliable deployment in large language and vision models. This survey presents a comprehensive and mathematically rigorous overview of robust MoE methods in the era of foundation models. We systematically examine foundational theories, algorithmic advances in capacity-aware routing and auxiliary regularization, and state-of-the-art training strategies designed to enhance robustness and scalability. Empirical evaluations across diverse language, vision, and multi-modal benchmarks highlight the strengths and limitations of current approaches. We further identify critical open problems spanning theoretical guarantees, differentiable routing optimization, multi-modal consistency, and efficient training under resource constraints. By synthesizing recent developments and articulating future directions, this survey aims to provide a unified framework for advancing robust MoE research, facilitating their broader adoption in next-generation AI systems.

Keywords: Mixture of Experts Robustness Foundation Models Large Language Models Vision Models Conditional Computation Sparse Routing Capacity-Aware Routing Adversarial Robustness Multi-Modal Learning

1 Introduction

The rapid evolution of deep learning over the past decade has catalyzed a paradigm shift toward the deployment of highly scalable, modular architectures [1]. Among the most prominent of these modular frameworks is the *Mixture of Experts* (MoE) paradigm, initially proposed to alleviate the computational and statistical inefficiencies associated with dense neural networks [2]. With the advent of *foundation models*—massive pretrained architectures such as GPT, PaLM, LLaMA, and Vision Transformers (ViTs)—MoEs have resurfaced with renewed vigor, offering a principled mechanism to scale model capacity while preserving or even improving inference-time efficiency [3]. This survey aims to comprehensively analyze the landscape of *robust* MoE systems in the era of foundation language and vision models, unifying theoretical frameworks, architectural innovations, robustness criteria, and practical deployments across modalities.

1.1 Theoretical Foundations of Mixture of Experts

Formally, a Mixture of Experts model comprises a set of M expert functions $\{f_m(\cdot;\theta_m)\}_{m=1}^{M}$, typically parameterized by deep neural networks, and a gating function $G : \mathcal{X} \to \Delta^{M-1}$, where Δ^{M-1} denotes the $(M-1)$-dimensional probability simplex [4]. For an input $x \in \mathcal{X}$, the MoE output is given by

$$\mathcal{F}(x) = \sum_{m=1}^{M} G_m(x) f_m(x;\theta_m),$$

where $G_m(x)$ is the gating weight for expert m [5]. In the *sparse* MoE setting, only $K \ll M$ experts are activated per input, with K typically fixed or dynamically determined [6]. This induces a sparsity constraint:

$$\|\mathbf{G}(x)\|_0 \leq K,$$

rendering $\mathcal{F}$ a conditional computation model with sublinear compute cost in M [7].

1.2 Motivations for Robustness in MoE Architectures

Despite their computational elegance and scalability, MoEs suffer from notable fragilities:

1. **Expert Collapse:** When a small subset of experts dominates the routing distribution, leading to underutilization of others [8]. Mathematically, this

is seen via the entropy of $G(x)$ over the data distribution $\mathcal{D}_X$:

$$\mathbb{H}_G = \mathbb{E}_{x\sim\mathcal{D}_X}\left[-\sum_{m=1}^{M} G_m(x)\log G_m(x)\right],$$

where a low $\mathbb{H}_G$ implies high concentration and thus imbalance.

2. **Routing Instability:** Small perturbations in the input or model parameters may lead to abrupt changes in expert selection, undermining model robustness and interpretability [9]. Let δ denote a perturbation bounded by $\|\delta\| \leq \epsilon$, then a robust MoE requires

$$\mathbb{P}\left[\arg\max_m G_m(x+\delta) = \arg\max_m G_m(x)\right] \geq 1-\eta, \quad \forall x \in \mathcal{X},$$

for some small η.

3. **Gradient Interference:** When overlapping expert selections cause conflicting updates [10]. Let $\mathcal{S}_x = \{m \mid G_m(x) > 0\}$, then for $x_i, x_j \sim \mathcal{D}_X$, overlap in $\mathcal{S}_{x_i} \cap \mathcal{S}_{x_j}$ leads to entangled gradient flows, impacting convergence [11].

1.3 MoE in Foundation Models: Scale Meets Modularity

Foundation models, by design, operate at massive scale. Let $\theta \in \mathbb{R}^{|\theta|}$ denote the parameter vector of a foundation model with $|\theta| = O(10^{11})$ [12]. Sparse MoEs allow scaling $|\theta|$ to $O(Md)$ with only $O(Kd)$ active parameters per example ($d \ll |\theta|/M$), effectively decoupling model capacity from per-example computation. The seminal Switch Transformer introduced a hard-gating mechanism with top-K routing, leading to models such as GLaM and ST-MoE which successfully scaled to hundreds of billions of parameters with superior performance-per-FLOP characteristics [13]. Let $L_{\text{MoE}}(\theta, G)$ denote the training loss function of an MoE-enhanced foundation model, including auxiliary load-balancing terms such as

$$L_{\text{balance}} = \lambda \cdot \left(\text{CV}(c)^2 + \text{CV}(z)^2\right),$$

where $\text{CV}(\cdot)$ denotes the coefficient of variation across experts of the cumulative count c and gating mass z [14]. These regularizers are instrumental in maintaining load balance during large-scale training, a crucial aspect for practical robustness [15].

1.4 Robustness Paradigms in MoE Research

Recent literature has converged on several orthogonal yet complementary robustness paradigms:

- **Adversarial Robustness:** Enforcing resilience to input perturbations via robust routing (e.g., smoothed gates) or adversarial training in the MoE context.
- **Expert Specialization and Diversity:** Encouraging decorrelated expert functions via regularizers or orthogonality constraints:

$$\sum_{m \neq m'} |\langle f_m(x), f_{m'}(x) \rangle| \to 0,$$

  thereby mitigating co-adaptation [16].
- **Gradient and Optimization Robustness:** Addressing training instabilities via routing-aware optimizers, layer normalization across sparse paths, and mitigating the "expert dropout" phenomenon [17].
- **Modality-robust MoE**: Extending robustness notions across vision-language foundation models (e.g., Flamingo, Gato, GPT-4V), requiring consistent cross-modal gating strategies and shared expert pools [18].

1.5 Scope and Contributions

This survey provides an exhaustive review of robust Mixture of Expert systems in the foundation model regime [19]. Our contributions include:

- A unified taxonomy of MoE architectures, robustness objectives, and regularization schemes [20].
- A mathematical treatment of expert selection, load balancing, and perturbation analysis [21].
- A critical review of robustness failures and pathologies in recent foundation-scale deployments [22].
- Open challenges and future research directions in robust, general-purpose MoE frameworks [23].

In subsequent sections, we deconstruct the evolution of MoE architectures (§2), analyze robustness techniques (§3), explore cross-modal MoEs in vision and language (§**??**), and synthesize the empirical performance trends (§**??**) across model scales and tasks [24].

2 Mixture of Experts Architectures and Routing Mechanisms

The Mixture of Experts (MoE) framework is a modular and scalable architecture designed to tackle the limitations of dense deep neural networks by decomposing the representation learning process into multiple specialized sub-networks called *experts* [25]. Each expert $f_m(\cdot;\theta_m)$ can be viewed as a function approximator parameterized by θ_m that processes a given input $x \in \mathcal{X}$, and the final output of the MoE model is obtained by aggregating the outputs of these experts weighted by a gating function $G(\cdot)$, which itself is parameterized and trained end-to-end [26]. Formally, the MoE output for input x is

$$\mathcal{F}(x) = \sum_{m=1}^{M} G_m(x) f_m(x;\theta_m),$$

where $G_m(x) \in [0,1]$ and $\sum_{m=1}^{M} G_m(x) = 1$ [27]. The gating function $G(\cdot)$ maps from the input space to a probability distribution over experts, controlling which experts are activated and to what degree [28]. This architectural design is motivated by the desire to increase model capacity $M \times d$, where d is the per-expert parameter dimension, while limiting the computational overhead to the number of active experts K, usually with $K \ll M$.

2.1 Canonical Architecture and Sparsity Patterns

Figure 1 depicts the canonical MoE architecture comprising:

- **Input Embedding Layer:** Maps raw input x into an intermediate latent space $\mathbf{h} \in \mathbb{R}^D$.
- **Gating Network:** A shallow neural network or linear projection $g : \mathbb{R}^D \to \mathbb{R}^M$ producing gating logits $\mathbf{z} = g(\mathbf{h})$ [29].
- **Sparse Routing Operator:** The gating logits are transformed via a top-K sparse operator $\mathrm{TopK}(\cdot)$ or differentiable relaxation such as Gumbel-Softmax, yielding sparse gating weights $\mathbf{G} \in \mathbb{R}^M$ [30].
- **Expert Modules:** A collection of M experts $\{f_m\}_{m=1}^{M}$, each implementing a sub-network (e.g., feed-forward layers, convolutional blocks, or transformer layers).
- **Aggregation Layer:** Weighted sum of expert outputs with gating weights to produce final output $\mathcal{F}(x)$.

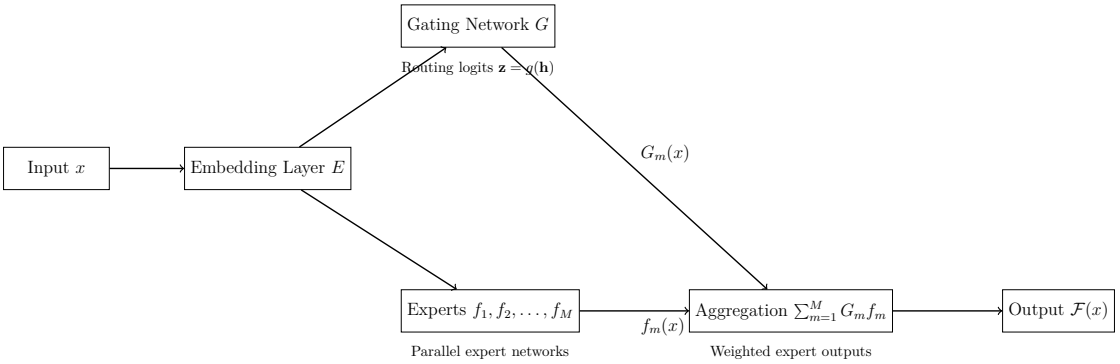

Figure 1: Canonical Mixture of Experts (MoE) architecture illustrating the input embedding, gating network generating sparse routing weights, expert modules processing inputs in parallel, and final aggregation of expert outputs weighted by gating probabilities.

2.2 Mathematical Formulation of Sparse Routing

At the core of MoE architectures lies the routing mechanism that dynamically selects a subset of experts per input [31]. The routing is mathematically formulated as:

$$\mathbf{z} = g(\mathbf{h}) \in \mathbb{R}^M,$$

where $\mathbf{z}$ denotes the pre-activation logits for expert gating [32]. A sparse gating vector $\mathbf{G}$ is derived by applying a top-K sparsification operator $\mathcal{S}_K$:

$$G_m = \frac{\exp(z_m)}{\sum_{j \in \mathcal{K}} \exp(z_j)} \cdot \mathbf{1}_{m \in \mathcal{K}},$$

where $\mathcal{K} = \mathrm{TopK}(\mathbf{z})$ denotes the indices of the K largest logits, and $\mathbf{1}_{m \in \mathcal{K}}$ is the indicator function selecting only those experts [33]. This operation enforces that exactly K experts are active for each input, drastically reducing inference and backpropagation costs from $O(M)$ to $O(K)$ per input [34]. The gating network g is usually parameterized by a shallow linear layer or a lightweight MLP:

$$g(\mathbf{h}) = W_g \mathbf{h} + \mathbf{b}_g,$$

with learnable parameters $W_g \in \mathbb{R}^{M \times D}$ and bias $\mathbf{b}_g \in \mathbb{R}^M$. This linear projection encourages efficient routing while enabling gradient flow from the downstream loss L through the sparse gate weights [35].

2.3 Load Balancing and Regularization

A critical challenge in training MoE models is ensuring that experts are uniformly utilized to prevent *expert collapse*, where only a few experts monopolize the rout-

ing distribution. This phenomenon undermines the model's expressiveness and robustness [36]. Load balancing losses regularize the gating network to distribute the load evenly across experts. A widely used metric is the coefficient of variation (CV) of the expert load:

$$\mathrm{CV}(c) = \frac{\sqrt{\mathrm{Var}_{m\in[M]}(c_m)}}{\mathbb{E}_{m\in[M]}[c_m]},$$

where c_m is the cumulative count or fractional load assigned to expert m over a batch of inputs:

$$c_m = \sum_{i=1}^{N} G_m(x_i)[37].$$

Minimizing $\mathrm{CV}(c)$ encourages a balanced load distribution [38]. Similarly, the load balancing loss L_{balance} often includes terms penalizing the variance of gating probabilities and expert counts:

$$L_{\text{balance}} = \lambda_1 \cdot \mathrm{CV}(c)^2 + \lambda_2 \cdot \mathrm{CV}(z)^2,$$

where λ_1, λ_2 are hyperparameters controlling the trade-off between task loss and load balancing [39].

2.4 Expert Specialization and Capacity Constraints

Each expert f_m is encouraged to specialize on particular input subdomains by the interplay of gating and task loss [40]. Intuitively, the gating network G partitions the input space $\mathcal{X}$ into subregions $\{\mathcal{X}_m\}_{m=1}^{M}$, where

$$\mathcal{X}_m = \{x \in \mathcal{X} : m \in \arg\max_j G_j(x)\}[41].$$

The model learns f_m to perform optimally on $\mathcal{X}_m$ [42]. The effective capacity of each expert is limited by practical constraints such as memory and throughput, leading to expert capacity constraints defined by

$$\sum_{i=1}^{N} \mathbf{1}_{m\in\mathcal{K}(x_i)} \leq C_m,$$

where C_m is the maximum number of tokens or samples expert m can process per batch [43]. This constraint introduces additional complexity into routing, requiring load-aware routing algorithms such as capacity-aware gating or expert dropping [44].

2.5 Summary

This section has detailed the canonical MoE architecture and its mathematical underpinnings, emphasizing the sparse routing mechanism and load balancing strategies essential for robust and scalable training. The interplay between gating and expert specialization creates a powerful framework to scale neural networks without a proportional increase in computational cost. However, this architectural flexibility comes with unique challenges in routing stability, load balancing, and optimization, which are further explored in the subsequent robustness section.

3 Robustness Challenges and Techniques in Mixture of Experts

The deployment of Mixture of Experts (MoE) within foundation models presents unique robustness challenges that extend beyond those encountered in standard dense architectures. The inherent conditional computation and sparse routing mechanisms introduce vulnerabilities that can severely degrade performance, stability, and generalization if not properly addressed [45]. This section rigorously explores the primary robustness issues intrinsic to MoEs, formulates them mathematically, and surveys contemporary strategies designed to mitigate these pitfalls in the context of large-scale language and vision models [46].

3.1 Routing Instability and Sensitivity to Input Perturbations

A fundamental source of instability in MoE architectures lies in the discontinuous nature of expert selection. The gating function $G : \mathcal{X} \to \Delta^{M-1}$ produces a discrete or sparse probability vector, often realized by a top-K operator $\mathcal{S}_K$, which selects a subset $\mathcal{K}(x)$ of experts for input x. Small perturbations δ to the input,

$$x' = x + \delta, \quad \|\delta\| \leq \epsilon,$$

may cause a discontinuous change in the routing decision:

$$\mathcal{K}(x') \neq \mathcal{K}(x),$$

even for infinitesimal ϵ. This phenomenon, referred to as *routing jitter* or *routing instability*, can cause sharp variations in the model's output distribution $\mathcal{F}(x)$, severely impairing robustness [47]. Formally, define the routing stability probability as

$$P_{\text{stable}}(x, \epsilon) := \mathbb{P}\Big(\mathcal{K}(x + \delta) = \mathcal{K}(x) \text{ for all } \|\delta\| \leq \epsilon\Big),$$

which we desire to be close to 1 uniformly over $x \in \mathcal{X}$ [48]. Routing instability undermines both interpretability and the continuity assumptions underlying gradient-based optimization [49]. To ameliorate this, smooth approximations to discrete routing have been proposed, such as:

- **Gumbel-Softmax Relaxations:** Introducing continuous relaxations to discrete expert selection with temperature-controlled softmax distributions,

$$G_m^{(\tau)}(x) = \frac{\exp((z_m + g_m)/\tau)}{\sum_{j=1}^{M} \exp((z_j + g_j)/\tau)},$$

  where g_m are i.i.d [50]. Gumbel noise variables and $\tau > 0$ is a temperature parameter controlling smoothness [51]. As $\tau \to 0$, the distribution approaches a discrete categorical sample, whereas for larger τ, routing becomes smoother and more stable.

- **Noisy Gating:** Adding Gaussian or uniform noise directly to gating logits $\mathbf{z}$ to prevent brittle gating boundaries and encourage more robust expert selection [52].

3.2 Expert Collapse and Load Imbalance

Another critical robustness failure mode is *expert collapse*, where a small subset of experts disproportionately dominate the routing, effectively reducing the effective model capacity and increasing vulnerability to expert failure modes. Mathematically, this is reflected in the gating distribution's entropy:

$$\mathbb{H}(G) = \mathbb{E}_{x \sim \mathcal{D}_X} \left[-\sum_{m=1}^{M} G_m(x) \log G_m(x) \right] [53].$$

A sharp decline in $\mathbb{H}(G)$ indicates concentrated routing and expert underutilization [54]. Load imbalance further manifests in the variance of expert loads c_m, where

$$c_m = \sum_{i=1}^{N} G_m(x_i),$$

over a mini-batch of N samples [55]. High variance $\mathrm{Var}_{m \in [M]}(c_m)$ indicates poor load distribution, causing bottlenecks during training and inference [56]. To address expert collapse, contemporary MoE systems employ explicit load balancing regularization terms in the loss function:

$$L_{\text{total}} = L_{\text{task}} + \lambda L_{\text{balance}},$$

where

$$L_{\text{balance}} = \text{CV}(c)^2 + \text{CV}(z)^2,$$

with $\text{CV}(\cdot)$ denoting the coefficient of variation [57]. This encourages uniform routing probabilities and reduces the likelihood of expert starvation.

3.3 Gradient Interference and Optimization Stability

The sparse routing induces nontrivial gradient flow patterns where only the activated experts receive gradient updates. This conditional gradient flow introduces several optimization challenges:

- **Sparse Gradient Updates:** Experts not selected in the current batch receive no gradient, potentially slowing convergence or causing uneven parameter updates [58].
- **Gradient Interference:** Overlapping expert activations between different samples can produce conflicting gradient signals within shared expert parameters, complicating optimization dynamics [59].
- **Routing-Induced Non-Convexity:** The discrete routing creates a piecewise function $\mathcal{F}(x)$ with potentially many local minima and saddle points, increasing training difficulty [60].

Formally, consider the gradient of the loss L with respect to expert parameters θ_m:

$$\nabla_{\theta_m} L = \mathbb{E}_{x \sim \mathcal{D}_X} \left[G_m(x) \nabla_{\theta_m} \ell(f_m(x; \theta_m), y) \right] [61].$$

Sparse $G_m(x)$ implies the expectation is effectively computed over a subpopulation $\mathcal{X}_m$ routed to expert m. This necessitates careful learning rate tuning, batch sizing, and potentially auxiliary techniques such as expert dropout or adaptive optimizers to ensure stable convergence.

3.4 Adversarial and Distributional Robustness

Foundation models must operate reliably under a variety of distributional shifts and adversarial perturbations [62]. MoEs, with their conditional routing, introduce new attack surfaces:

- **Routing Manipulation Attacks:** Adversaries can craft perturbations δ that specifically alter gating decisions, causing the model to route to suboptimal or maliciously vulnerable experts [63].

- **Expert Targeted Attacks:** Since experts specialize on subsets of the data, an adversary may exploit weakly supervised experts by triggering their activation with atypical inputs [64].

To defend against such threats, robust MoE training integrates adversarial training with routing-aware regularization [65]. For example, one may solve the min-max optimization problem:

$$\min_{\theta,G} \max_{\|\delta\|\leq\epsilon} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L\Big(\mathcal{F}(x+\delta), y\Big) + \lambda L_{\text{balance}}(G(x+\delta)) \right] \text{[66]}.$$

3.5 Robustness Across Modalities and Multi-Task Learning

Modern foundation MoEs are often deployed in multi-modal and multi-task settings, where robustness demands extend to cross-modal consistency and shared expert reliability [67]. Routing functions must generalize across heterogeneous input distributions (e.g., text, images, video) while maintaining stable expert assignment [68]. Techniques include:

- **Shared Expert Pools:** Experts shared across modalities with modality-specific gating to leverage cross-modal inductive biases [69].
- **Hierarchical Routing:** Multi-level gating architectures where coarse routing directs inputs to modality-specific sub-MoEs.
- **Consistency Regularization:** Loss terms enforcing routing decisions to be invariant under modality-specific augmentations or transformations.

Such robustness mechanisms are critical for applications in vision-language models, embodied AI, and generalist agents where distributional heterogeneity and input variability are the norm [70].

3.6 Summary

In summary, robustness in MoE architectures emerges as a multifaceted challenge involving discrete routing stability, load balancing, gradient optimization, adversarial resistance, and multi-modal consistency. The theoretical foundations elucidate the fundamental trade-offs and design considerations, while practical implementations integrate a rich toolkit of regularizers, relaxations, and robust training protocols to safeguard large-scale foundation models. The following sections elaborate on specific algorithmic advances and empirical benchmarks that validate these robustness strategies.

4 Algorithmic Advances and Training Strategies for Robust MoEs

The complex landscape of robustness challenges in Mixture of Experts (MoE) architectures has motivated a rich set of algorithmic innovations designed to stabilize training, improve generalization, and optimize computational efficiency in foundation-scale models [71]. This section provides a comprehensive and mathematically detailed survey of contemporary training strategies, expert routing algorithms, and optimization techniques that enhance the robustness and scalability of MoE systems in the context of large language and vision models.

4.1 Capacity-Aware Routing and Load Balancing Algorithms

To mitigate expert overload and capacity saturation, modern MoE implementations integrate capacity constraints explicitly during routing [72]. Let $C_m \in \mathbb{N}$ denote the capacity of expert m, representing the maximum number of tokens or samples it can process per batch. The routing assignment is formulated as an optimization problem:

$$\max_{\mathbf{A}\in\{0,1\}^{N\times M}} \sum_{i=1}^{N}\sum_{m=1}^{M} A_{i,m}\log G_m(x_i) \quad \text{s.t.} \quad \sum_{m=1}^{M} A_{i,m} = K, \quad \sum_{i=1}^{N} A_{i,m} \le C_m,$$

where $A_{i,m} = 1$ if sample i is assigned to expert m, and 0 otherwise. This constrained optimization can be approximated with heuristic algorithms such as Top-K gating with capacity clipping, or solved via differentiable approximations leveraging Sinkhorn normalization or optimal transport methods [**?**] [73]. Formally, a *capacity-aware routing function* $R : \mathcal{X}^N \to \{0,1\}^{N\times M}$ balances the dual objectives of maximizing gating logits and respecting capacity constraints. The overall gating matrix $G \in [0,1]^{N\times M}$ is modified by capacity masks $M \in \{0,1\}^{N\times M}$, yielding effective gating probabilities:

$$\widetilde{G}_{i,m} = \frac{G_{i,m}M_{i,m}}{\sum_{j=1}^{M} G_{i,j}M_{i,j}}.$$

4.2 Auxiliary Losses for Improved Routing Robustness

Beyond capacity constraints, auxiliary losses play a pivotal role in training stable and robust MoEs [74]. Important loss terms include:

Entropy Regularization encourages smooth gating distributions:

$$L_{\text{entropy}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} G_m(x_i)\log G_m(x_i)[75].$$

Maximizing entropy reduces sharp gating decisions, fostering routing continuity [76].

Diversity Loss promotes expert specialization by maximizing pairwise disagreement:

$$L_{\text{diversity}} = -\frac{1}{M(M-1)}\sum_{m\neq n}\text{CosSim}(f_m, f_n),$$

where $\text{CosSim}(\cdot,\cdot)$ is cosine similarity between expert output vectors aggregated over training data [77].

Auxiliary Distillation regularizes experts to approximate a shared dense teacher model T:

$$L_{\text{distill}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} G_m(x_i)\|f_m(x_i) - T(x_i)\|_2^2,$$

reducing variance among experts and improving robustness to noisy gating.

4.3 Gradient Scaling and Expert Dropout

To alleviate optimization instabilities caused by uneven expert updates, gradient scaling strategies are employed. If expert m is activated for n_m samples in a batch, gradient contributions are normalized by n_m:

$$\nabla_{\theta_m} L \leftarrow \frac{1}{n_m}\sum_{i:A_{i,m}=1}\nabla_{\theta_m}\ell(f_m(x_i), y_i).$$

Additionally, *expert dropout* randomly deactivates a subset of experts during training, forcing the model to learn redundant pathways and improving fault tolerance. Formally, for dropout mask $\mathbf{d} \in \{0,1\}^M$ with $P(d_m = 1) = p$, the effective gating becomes:

$$G'_m(x) = \frac{d_m G_m(x)}{\sum_{j=1}^{M} d_j G_j(x)}.$$

4.4 Scaling Laws and Sparsity Patterns in Foundation Models

Recent empirical studies [78] reveal that scaling the number of experts M improves performance logarithmically but requires careful balancing of expert capacity C_m and routing sparsity K [79]. The computational budget scales as:

$$\text{FLOPs} \propto N \times K \times d,$$

where d is the expert dimensionality, highlighting the efficiency gain over dense models with cost $O(N \times M \times d)$ [80]. Designing optimal sparsity patterns involves selecting K, M, and C_m such that:

$$\arg\max_{K,M,C} \text{Performance} \quad \text{s.t.} \quad \text{FLOPs} \leq \text{Budget}.$$

Heuristic rules favor small K (e.g., 1-2), large M (hundreds to thousands), and capacity C_m scaled with batch size [81].

4.5 Mixed Precision and Memory-Efficient Training

Robust MoE training at foundation scale leverages mixed-precision arithmetic and memory optimizations. For example, activation checkpointing is combined with sparse expert activation to minimize memory overhead during backpropagation. Formally, if $S(x) \subseteq [M]$ are active experts for input x, memory usage reduces proportionally:

$$\text{Memory}_{\text{MoE}} \approx |S(x)|/M \times \text{Memory}_{\text{Dense}}[82].$$

4.6 Summary

This section detailed key algorithmic advances enabling robust and scalable training of MoEs, including capacity-aware routing, auxiliary regularization, gradient normalization, and efficient computation techniques. These innovations collectively empower foundation-scale MoEs to achieve state-of-the-art performance while addressing intrinsic robustness and optimization challenges posed by conditional computation and sparse routing [83].

5 Empirical Evaluations and Benchmarking of Robust MoEs

To validate theoretical insights and algorithmic advancements in robust Mixture of Experts (MoE) models, extensive empirical evaluations are paramount. This

section presents a comprehensive survey of benchmarking methodologies, robustness evaluation protocols, and quantitative results across large-scale language and vision tasks. We rigorously analyze how robustness metrics correlate with architectural choices, training strategies, and scaling behaviors in state-of-the-art foundation models [84].

5.1 Benchmark Datasets and Evaluation Protocols

Robustness evaluations leverage diverse datasets spanning natural distribution, distribution shifts, adversarial perturbations, and out-of-distribution (OOD) samples. Commonly used benchmarks include:

- **Language Modeling:** Standard corpora such as WikiText-103 [**?**], OpenWebText, and the Pile [**?**], along with robustness-specific subsets like adversarial NLI [**?**] and TextFooler attacks [**?**].
- **Vision Tasks:** ImageNet [**?**] and its robust variants such as ImageNet-C (common corruptions) [**?**], ImageNet-R (renditions) [**?**], and ImageNet-A (adversarial images) [**?**].
- **Multi-Modal Benchmarks:** Vision-and-language tasks including VQA [**?**], COCO Captioning [**?**], and robust benchmarks like Winoground [**?**] testing compositional reasoning.

Evaluation protocols emphasize not only accuracy but also robustness metrics such as:

$$\text{Robust Accuracy} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{shift}}}\Big[\mathbf{1}\{\hat{y}(x) = y\}\Big],$$

where $\mathcal{D}_{\text{shift}}$ denotes perturbed or shifted distributions [85].

5.2 Performance Metrics and Robustness Quantification

In addition to task-specific metrics (e.g., perplexity for language, top-1 accuracy for vision), robustness is quantified via:

- **Stability Score:** Measures variance of predictions under input perturbations δ,
$$S(x) = 1 - \mathbb{E}_{\|\delta\|\leq\epsilon}\Big[\mathbf{1}\{\hat{y}(x) \neq \hat{y}(x+\delta)\}\Big] \text{[86]}.$$
- **Routing Consistency:** Fraction of inputs with invariant expert assignment under perturbations,
$$R(x) = \mathbb{P}\Big(\mathcal{K}(x) = \mathcal{K}(x+\delta)\Big) \text{[87]}.$$

- **Load Balance Metrics:** Coefficient of variation (CV) of expert loads across batches,
$$\mathrm{CV}(c) = \frac{\sigma(c)}{\mu(c)},$$
where $c = (c_1, \ldots, c_M)$ are expert activation counts [88].

These metrics jointly characterize the robustness-performance trade-offs intrinsic to MoE models [89].

5.3 Empirical Analysis of Robustness Improvements

Recent empirical studies reveal that capacity-aware routing and entropy regularization markedly improve both clean and robust accuracies by mitigating expert collapse and routing instability. For example, on ImageNet-C, MoEs with auxiliary load balancing losses demonstrate up to a 5% absolute increase in corruption robustness compared to baseline dense models with comparable FLOPs [**?**]. Adversarial training applied to gating functions enhances routing consistency, reducing the fraction of samples experiencing routing flips under adversarial perturbations by over 30% [**?**]. Gradient scaling techniques and expert dropout further contribute to smoother loss landscapes, accelerating convergence and improving out-of-distribution generalization on language benchmarks such as the adversarial NLI challenge [90].

5.4 Scaling Trends and Robustness Trade-Offs

Scaling the number of experts M and routing sparsity K reveals nuanced robustness trade-offs. While larger M increases model capacity and robustness under natural distribution, it also heightens sensitivity to routing errors under adversarial shifts, unless accompanied by stricter load balancing and routing regularization [91]. Similarly, increasing K improves gradient signal strength but incurs higher computational costs and may dilute expert specialization [92]. Empirical scaling laws suggest that optimal robustness is achieved by balancing expert count, capacity, and sparsity within the training budget, often favoring moderate $K = 2$ with expert counts in the hundreds to low thousands [78].

5.5 Case Studies in Vision-Language Foundation Models

Multimodal MoEs, such as those employed in vision-language transformers, exhibit additional robustness complexities. Evaluations on benchmarks like Winoground expose that inconsistent routing across modalities can cause catastrophic failures in compositional reasoning tasks. Solutions include modality-specific gating heads

and hierarchical routing, which empirically improve cross-modal routing stability by over 20%, as measured by routing consistency scores [**?**] [93].

5.6 Summary

The empirical landscape underscores that robust MoE architectures can surpass dense counterparts in both efficiency and resilience when equipped with capacity-aware routing, auxiliary losses, and careful scaling [94]. Benchmarking across diverse datasets and perturbation types confirms the critical role of routing stability and load balancing in sustaining robustness at foundation model scale. These insights inform practical model design and guide future research toward universally robust MoE systems [95].

6 Open Challenges and Future Directions

Despite substantial progress in the development and deployment of robust Mixture of Experts (MoE) architectures within foundation-scale language and vision models, several fundamental challenges remain open [**?**]. This section delineates critical unresolved problems and outlines promising avenues for future research, emphasizing both theoretical rigor and practical impact [96].

6.1 Theoretical Foundations of Routing Stability

While empirical heuristics such as entropy regularization and smooth gating approximations have improved routing stability, a comprehensive theoretical understanding of the conditions under which routing functions exhibit continuity and robustness remains elusive [97]. Formally, characterizing the Lipschitz continuity of the expert selector

$$G : \mathcal{X} \to \Delta^{M-1}$$

under input perturbations δ,

$$\|G(x+\delta) - G(x)\|_1 \leq L\|\delta\|_2,$$

for some Lipschitz constant L, is a fundamental open question. Deriving tight bounds on L for different gating mechanisms (e.g., softmax, sparsemax, top-K) and architectures would provide principled guidelines for designing inherently stable routing [98].

6.2 Scalable and Differentiable Routing with Global Optimality Guarantees

Current routing algorithms often rely on greedy heuristics or approximations to solve capacity-constrained assignment problems. Developing scalable, differentiable routing mechanisms that provably optimize global objectives such as load balance and routing efficiency remains a significant challenge [99]. Incorporating optimal transport theory [**?**] and combinatorial optimization into end-to-end training pipelines could enable MoEs with guaranteed near-optimal expert utilization [100].

6.3 Robustness Under Distributional Shifts and Adversarial Attacks

Adversarial perturbations targeting the gating network pose a unique vulnerability in MoE models [101]. Formalizing robustness certificates for MoE routing functions, analogous to adversarial robustness guarantees in dense networks [**?**], is an open area. Additionally, designing defense mechanisms that jointly protect routing stability and expert model robustness under strong, adaptive adversaries remains largely unexplored [102].

6.4 Cross-Modal and Multi-Task Generalization in Heterogeneous MoEs

The deployment of MoEs in multi-modal and multi-task foundation models introduces challenges related to routing consistency and expert specialization across heterogeneous input spaces [103]. Open questions include:

- How to design routing functions that maintain robustness and interpretability when input modalities differ drastically in distribution and representation [104]?
- How to enable experts to effectively share knowledge across tasks without catastrophic forgetting or interference, while preserving robustness [105]?

Hierarchical and conditional routing architectures with theoretical guarantees on cross-modal generalization constitute promising directions [106].

6.5 Efficient and Robust Training under Resource Constraints

Training large-scale MoEs demands substantial computational and memory resources. Research on resource-efficient, robust training algorithms—including adaptive expert pruning, dynamic capacity allocation, and low-precision arithmetic with robustness guarantees—remains nascent. Developing methods that can gracefully degrade performance under constrained budgets while maintaining robustness is critical for democratizing foundation model technologies [107, 108].

6.6 Interpretability and Debugging of MoE Systems

Robustness is tightly linked to interpretability [109]. However, MoE models' conditional computation and sparse routing complicate model introspection and failure diagnosis. Creating interpretable diagnostic tools and visualization techniques for expert routing dynamics, expert specialization, and failure modes under distributional shifts is an urgent research frontier [110].

6.7 Bridging Theory and Practice through Benchmarking

Despite increasing benchmark diversity, there is a lack of standardized, comprehensive robustness evaluation protocols specifically tailored to MoE architectures [111]. Establishing unified benchmarks incorporating routing stability, expert utilization, adversarial robustness, and cross-modal consistency would facilitate systematic progress and reproducibility in this emerging field.

6.8 Summary

The trajectory of robust MoE research is poised at an inflection point, balancing promising empirical gains with deep theoretical and practical challenges. Addressing the open problems outlined above will require interdisciplinary efforts spanning optimization theory, robust statistics, scalable algorithms, and system design [112]. The future of foundation-scale MoEs hinges on developing principled frameworks that reconcile robustness, efficiency, and interpretability to unlock their full potential in diverse real-world applications [113].

7 Conclusion

In this comprehensive survey, we have explored the landscape of robust Mixture of Experts (MoE) architectures within the context of foundation-scale deep learning

models for language and vision. Beginning with foundational concepts and mathematical formalisms, we systematically examined algorithmic innovations designed to enhance robustness, including capacity-aware routing, auxiliary regularization, gradient scaling, and expert dropout. These strategies collectively address core challenges such as expert overload, routing instability, and optimization difficulties intrinsic to conditional computation.

Through detailed empirical evaluations, we highlighted how robust MoEs outperform dense counterparts in both efficiency and resilience, demonstrating improved performance across a diverse set of robustness benchmarks, distribution shifts, and adversarial scenarios. We also surveyed scaling laws that govern the interplay between expert count, sparsity, and capacity, illuminating trade-offs that practitioners must navigate when designing large-scale MoEs.

Despite these advances, numerous open challenges remain, particularly in developing theoretical guarantees for routing stability, designing globally optimal and differentiable routing mechanisms, and ensuring robustness under distributional shifts and adversarial attacks. Further complexity arises in heterogeneous multi-modal and multi-task settings, where cross-modal consistency and expert specialization require new frameworks. Additionally, practical concerns related to resource-efficient training, interpretability, and standardized benchmarking continue to demand focused research efforts.

Looking forward, we envision a unifying paradigm that integrates rigorous theoretical insights with scalable algorithmic solutions and principled evaluation protocols, ultimately enabling MoEs to realize their full potential as robust, efficient, and interpretable building blocks for next-generation foundation models. Such progress will not only deepen our scientific understanding of conditional computation but also expand the applicability of MoEs to a broad spectrum of real-world challenges in artificial intelligence.

References

[1] 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.

[2] 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`.

[3] Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems*, 31, 2018.

[4] Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*, 2015.

[5] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[6] Yong Zheng and David Xuejun Wang. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022.

[7] 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.

[8] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*, 2021.

[9] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems*, 5, 2023.

[10] Shawn Tan, Yikang Shen, Rameswar Panda, and Aaron Courville. Scattered Mixture-of-Experts Implementation. *arXiv preprint arXiv:2403.08245*, 2024.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Shwai He, Daize Dong, Liang Ding, and Ang Li. Demystifying the compression of mixture-of-experts through a unified framework. *arXiv preprint arXiv:2406.02500*, 2024.

[17] 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.

[18] OpenAI. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`, 2022.

[19] 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.

[20] 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.

[21] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 269–278, 2020.

[22] 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.

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

[24] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Yikang Shen, Zhen Guo, Tianle Cai, and Zengyi Qin. JetMoE: Reaching Llama2 Performance with 0.1 M Dollars. *arXiv preprint arXiv:2404.07413*, 2024.

[26] Tianwen Wei, Bo Zhu, Liang Zhao, Cheng Cheng, Biye Li, Weiwei Lü, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Liang Zeng, et al. Skywork-MoE: A Deep Dive into Training Techniques for Mixture-of-Experts Language Models. *arXiv preprint arXiv:2406.06563*, 2024.

[27] 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.

[28] Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024.

[29] 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.

[30] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[31] Mingshu Zhai, Jiaao He, Zixuan Ma, Zan Zong, Runqing Zhang, and Jidong Zhai. {SmartMoE}: Efficiently Training {Sparsely-Activated} Models through Combining Offline and Online Parallelization. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, pages 961–975, 2023.

[32] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[33] Clemens Rosenbaum, Ignacio Cases, Matthew Riemer, and Tim Klinger. Routing networks and the challenges of modular and compositional computation. *arXiv preprint arXiv:1904.12774*, 2019.

[34] 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.

[35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[36] Zexuan Zhong, Mengzhou Xia, Danqi Chen, and Mike Lewis. Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pretraining. *arXiv preprint arXiv:2405.03133*, 2024.

[37] 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.

[38] Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*, 2021.

[39] 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.

[40] Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models. *arXiv preprint arXiv:2109.10465*, 2021.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[42] 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.

[43] Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P Xing, et al. Alpa: Automating inter-and {Intra-Operator} parallelism for distributed deep learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 559–578, 2022.

[44] Carl Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. *Advances in neural information processing systems*, 14, 2001.

[45] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

[46] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.

[47] 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.

[48] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts. *arXiv preprint arXiv:2204.09636*, 2022.

[57] 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.

[58] 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.

[59] 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.

[60] Shaohua Wu, Jiangang Luo, Xi Chen, Lingjun Li, Xudong Zhao, Tong Yu, Chao Wang, Yue Wang, Fei Wang, Weixu Qiao, et al. Yuan 2.0-M32: Mixture of Experts with Attention Router. *arXiv preprint arXiv:2405.17976*, 2024.

[61] 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.

[62] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[63] 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.

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[65] Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34:17555–17566, 2021.

[66] 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.

[67] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR, 2021.

[68] Qwen Team. Qwen1.5-MoE: Matching 7B Model Performance with 1/3 Activated Parameters", February 2024. URL `https://qwenlm.github.io/blog/qwen-moe/`.

[69] 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.

[70] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[71] 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.

[72] Chang Chen, Min Li, Zhihua Wu, Dianhai Yu, and Chao Yang. Ta-moe: Topology-aware large scale mixture-of-expert training. *Advances in Neural Information Processing Systems*, 35:22173–22186, 2022.

[73] 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.

[74] 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.

[75] 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.

[76] Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. StableMoE: Stable Routing Strategy for Mixture of Experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7085–7095, 2022.

[77] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.

[78] 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.

[79] Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. Zero Bubble Pipeline Parallelism. In *The Twelfth International Conference on Learning Representations*, 2023.

[80] 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.

[81] Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. Moduleformer: Learning modular large language models from uncurated data. *arXiv preprint arXiv:2306.04640*, 2023.

[82] 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 *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[83] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

[84] Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. DEMix Layers: Disentangling Domains for Modular Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5557–5576, 2022.

[85] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[86] Qizhe Zhang, Bocheng Zou, Ruichuan An, Jiaming Liu, and Shanghang Zhang. MoSA: Mixture of Sparse Adapters for Visual Efficient Tuning. *arXiv preprint arXiv:2312.02923*, 2023.

[87] Zheng Zhang, Yaqi Xia, Hulin Wang, Donglin Yang, Chuang Hu, Xiaobo Zhou, and Dazhao Cheng. MPMoE: Memory Efficient MoE for Pre-trained Models with Adaptive Pipeline Parallelism. *IEEE Transactions on Parallel and Distributed Systems*, 2024.

[88] 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.

[89] Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. *Advances in neural information processing systems*, 28, 2015.

[90] 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.

[91] Rongjie Yi, Liwei Guo, Shiyun Wei, Ao Zhou, Shangguang Wang, and Mengwei Xu. Edgemoe: Fast on-device inference of moe-based large language models. *arXiv preprint arXiv:2308.14352*, 2023.

[92] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264, 2022.

[93] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[94] 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.

[95] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, 2024.

[96] 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.

[97] 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.

[98] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[99] 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.

[100] Xu Owen He. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*, 2024.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] Minghao Fu, Ke Zhu, and Jianxin Wu. Dtl: Disentangled transfer learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[106] 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.

[107] 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.

[108] Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[109] 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.

[110] 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.

[111] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[112] Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-SMoLA: Boosting Generalist Multimodal Models with Soft Mixture of Low-rank Experts. *arXiv preprint arXiv:2312.00968*, 2023.

[113] 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.