Novelty Assessment Report

Paper: Coupling Experts and Routers in Mixture-of-Experts via an Auxiliary Loss

PDF URL: https://openreview.net/pdf?id=MpeyjgWbKt

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Traditional Mixture-of-Experts (MoE) models lack explicit constraints to ensure the router's decisions align well with the experts' capabilities, which ultimately limits model performance. To address this, we propose expert-router coupling loss (ERC loss), a lightweight auxiliary loss that couples expert capabilities and the router's decisions. We treat each row of the router matrix as a cluster center for the tokens assigned to a particular expert. From these centers, we create proxy tokens by applying a perturbation with noise. Using these proxy tokens, the ERC loss forces the router and experts to satisfy two constraints: (1) each expert exhibits higher activation for its corresponding proxy token than for any other proxy token, and (2) each proxy token elicits stronger activation in its designated expert than in any other expert. This optimization leads to two key effects: each row of the router matrix is an accurate representation of its expert's capabilities, while each expert develops expertise that closely match the tokens routed to it. Our experiments involve pre-training multiple 3B-parameter MoE-LLMs on trillions of tokens in total, providing detailed evidence of the ERC loss's effectiveness. Additionally, the ERC loss offers flexible control and quantitative tracking of expert specialization levels during training, providing many valuable insights into MoEs.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Aligning Router Decisions with Expert Capabilities in Mixture-of-Experts Models**

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

Taxonomy Overview

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

- **Router-Expert Alignment Mechanisms**
- **Router Design and Optimization Strategies**
- **Expert Specialization and Utilization**
- **Cross-Expert Routing and Coordination**
- **System-Level Optimization for MoE Inference**
- **Domain-Specific MoE Applications**
- **MoE Training and Optimization Foundations**
- **Security and Robustness in MoE**

Complete Taxonomy Tree

- Aligning Router Decisions with Expert Capabilities in Mixture-of-Experts Models Survey Taxonomy
- Router-Expert Alignment Mechanisms
  - Auxiliary Loss-Based Alignment ★ (3 papers)
  - [0] Coupling Experts and Routers in Mixture-of-Experts via an Auxiliary Loss (Anon et al., 2026) View paper
  - [7] Advancing Expert Specialization for Better MoE (Hongcan Guo, 2025) View paper
  - [25] On the representation collapse of sparse mixture of experts (Chi, 2022) View paper
  - Architectural Coupling Mechanisms (2 papers)
  - [44] DRAMoE: Boosting Adversarial Robustness with Adversarial Training and Adaptive Mixture of Experts (Y Fu, 2025) View paper
  - [45] Tight Clusters Make Specialized Experts (Stefan K. Nielsen, 2025) View paper
  - Representation and Manifold Alignment (2 papers)
  - [14] Training mixture-of-experts: A focus on expert-token matching (F Vesaghati, 2024) View paper
  - [30] Routing Manifold Alignment Improves Generalization of Mixture-of-Experts LLMs (Zhongyang Li, 2025) View paper
- Router Design and Optimization Strategies
  - Dynamic and Adaptive Routing (4 papers)
  - [2] Mixture-of-experts with expert choice routing (Zhou, 2022) View paper
  - [22] Rewiring Experts on the Fly: Continuous Rerouting for Better Online Adaptation in Mixture-of-Expert models (Su, 2025) View paper
  - [28] DA-MoE: Towards Dynamic Expert Allocation for Mixture-of-Experts Models (Maryam Akhavan Aghdam, 2024) View paper
  - [46] Efficient Routing in Sparse Mixture-of-Experts (Masoumeh Zareapoor, 2024) View paper
  - Hierarchical and Sequential Routing (3 papers)
  - [6] Novel token-level recurrent routing for enhanced mixture-of-experts performance (Ethan Pedicir, 2024) View paper
  - [23] Chain-of-Experts: Unlocking the Communication Power of Mixture-of-Experts Models (Wang Zihan, 2025) View paper
  - [43] Layerwise Recurrent Router for Mixture-of-Experts (Qiu, 2024) View paper
  - Domain-Specific Routing (5 papers)
  - [5] Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing (Zhou Hao, 2025) View paper

  - [10] Routing experts: Learning to route dynamic experts in multi-modal large language models (Wu Qiong, 2024) View paper
  - [26] Q-moe: Connector for mllms with text-driven routing (Hanzi Wang, 2024) View paper
  - [27] Multilingual Routing in Mixture-of-Experts (Bandarkar, 2025) View paper
  - [37] Sparse moe with language guided routing for multilingual machine translation (X Zhao, 2024) View paper
  - Inference-Time Routing Optimization (1 papers)
  - [24] From Score Distributions to Balance: Plug-and-Play Mixture-of-Experts Routing (Shahout Rana, 2025) View paper
- Expert Specialization and Utilization
  - Expert Specialization Analysis (4 papers)
  - [4] A closer look into mixture-of-experts in large language models (Lo, 2025) View paper
  - [11] Decoding Knowledge Attribution in Mixture-of-Experts: A Framework of Basic-Refinement Collaboration and Efficiency Analysis (Li, 2025) View paper
  - [19] Not All Models Suit Expert Offloading: On Local Routing Consistency of Mixture-of-Expert Models (Wang Si-yuan, 2025) View paper
  - [31] Part-Of-Speech Sensitivity of Routers in Mixture of Experts Models (Antoine Elie, 2024) View paper
  - Expert Specialization Enhancement (2 papers)
  - [42] How Do Shared Experts Dynamically Adapt to Routing Constraints in Mixture-of-Experts? (H Li, 2025) View paper
  - [47] Improving Expert Specialization in Mixture of Experts (Krishnamurthy, 2023) View paper
  - Expert Reuse and Upcycling (3 papers)
  - [9] Upcycling Instruction Tuning from Dense to Mixture-of-Experts via Parameter Merging (Hui, 2025) View paper
  - [36] : Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design (R Cai, 2024) View paper
  - [38] Router Upcycling: Leveraging Mixture-of-Routers in Mixture-of-Experts Upcycling (Ran Junfeng, 2025) View paper
- Cross-Expert Routing and Coordination
  - Multi-Model Expert Routing (3 papers)
  - [8] Learning to route among specialized experts for zero-shot generalization (Muqeeth, 2024) View paper
  - [12] ROUTERRETRIEVER: Routing over a Mixture of Expert Embedding Models (Hyunji Lee, 2025) View paper
  - [20] Symbolic Mixture-of-Experts: Adaptive Skill-based Routing for Heterogeneous Reasoning (Yun, 2025) View paper
  - Task-Guided Expert Selection (2 papers)
  - [32] Mixture of experts for network optimization: a large language model-enabled approach (DU Hongyang, 2024) View paper
  - [40] AT-MoE: Adaptive Task-planning Mixture of Experts via LoRA Approach (Li Xurui, 2024) View paper
- System-Level Optimization for MoE Inference
  - Expert Offloading and Memory Management (3 papers)
  - [16] Enabling MoE on the Edge via Importance-Driven Expert Scheduling (Zhu Guoying, 2025) View paper
  - [21] ExpertFlow: Adaptive Expert Scheduling and Memory Coordination for Efficient MoE Inference (Zixu Shen, 2025) View paper
  - [50] ExpertFlow: Optimized Expert Activation and Token Allocation for Efficient Mixture-of-Experts Inference (He, 2024) View paper
  - System Architecture and Hardware Co-Design (3 papers)
  - [3] A survey on inference optimization techniques for mixture of experts models (Liu Jiaâ□□Cheng, 2024) View paper
  - [13] MoE at Scale: From Modular Design to Deployment in Large-Scale Machine Learning Systems (Ahmed Al-Hassan, 2025) View paper
  - [48] Towards Using Partitioned GPU Virtual Functions for Mixture of Experts (Vinay Chander, 2024) View paper
- Domain-Specific MoE Applications
  - Vision and Multimodal MoE (3 papers)
  - [34] Vimoe: An empirical study of designing vision mixture-of-experts (Han Xumeng, 2024) View paper
  - [35] GM-MoE: Low-Light Enhancement with Gated-Mechanism Mixture-of-Experts (Liao Minwen, 2025) View paper
  - [41] Steer-MoE: Efficient Audio-Language Alignment with a Mixture-of-Experts Steering Module (Feng, 2025) View paper
  - Specialized Task MoE (3 papers)
  - [18] Frequency-Augmented Mixture-of-Heterogeneous-Experts Framework for Sequential Recommendation (Junjie Zhang, 2025) View paper
  - [39] OrdMoE: Preference Alignment via Hierarchical Expert Group Ranking in Multimodal Mixture-of-Experts LLMs (Yuting Gao, 2025) View paper
  - [49] What really matters for person re-identification? A Mixture-of-Experts Framework for Semantic Attribute Importance (Athena Psalta, 2025) View paper
- MoE Training and Optimization Foundations
  - Routing Mechanism Foundations (2 papers)
  - [1] On the benefits of learning to route in mixture-of-experts models (Nishanth Dikkala, 2023) View paper
  - [15] Mixture of Routers (Zhang Jia-Chen, 2025) View paper
  - Comprehensive MoE Surveys (2 papers)
  - [29] The evolution of moe: A survey from basics to breakthroughs (A Vats, 2024) View paper
  - [33] The evolution of mixture of experts: A survey from basics to breakthroughs (Arpita Vats, 2024) View paper
- Security and Robustness in MoE (1 papers)
  - [17] BadMoE: Backdooring Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts (Wang Qingyue, 2025) View paper

Narrative

Core task: Aligning router decisions with expert capabilities in mixture-of-experts models. The field has evolved around a central challenge—ensuring that routing mechanisms effectively match tokens or inputs to the most suitable experts within sparse MoE architectures. The taxonomy reflects this through several major branches: Router-Expert Alignment Mechanisms explores techniques such as auxiliary losses and coupling strategies to improve coordination between routers and experts (e.g., Coupling Experts Routers[0], Advancing Expert Specialization[7]); Router Design and Optimization Strategies addresses architectural choices like expert-choice routing (Expert Choice Routing[2]) and dynamic allocation schemes; Expert Specialization and Utilization examines how experts develop distinct capabilities and how to encourage meaningful differentiation; Cross-Expert Routing and Coordination investigates multi-expert collaboration and sequential routing patterns (Chain of Experts[23]); System-Level Optimization for MoE Inference focuses on computational efficiency and scheduling (MoE Inference Optimization[3], Importance Driven Scheduling[16]); Domain-Specific MoE Applications tailors routing to particular modalities or tasks (Multilingual Language Priors[5], Vision MoE Design[34]); MoE Training and

Optimization Foundations covers core training dynamics and load balancing; and Security and Robustness in MoE addresses vulnerabilities like backdoor attacks (BadMoE Backdooring[17]).

Recent work has intensified around preventing representation collapse and ensuring that auxiliary objectives genuinely promote expert diversity without undermining task performance—a tension visible in studies like Representation Collapse Sparse[25] and Closer Look MoE[4]. The original paper, Coupling Experts Routers[0], sits squarely within the Router-Expert Alignment Mechanisms branch, specifically targeting auxiliary loss-based alignment. It shares thematic ground with Advancing Expert Specialization[7], which also emphasizes tighter coupling between routing decisions and expert competencies, but differs in its focus on explicit loss formulations that directly penalize misalignment. Compared to works like Benefits Learning Route[1] or Token Recurrent Routing[6], which explore alternative routing paradigms or temporal dependencies, Coupling Experts Routers[0] prioritizes a more direct supervisory signal to guide routers toward experts' learned strengths, addressing a persistent challenge in balancing load distribution with specialization quality.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Advancing Expert Specialization for Better MoE

Authors: Hongcan Guo, Nan, Guoshun, Haolang Lu, Guoshun Nan, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Mixture-of-Experts (MoE) models enable efficient scaling of large language models (LLMs) by activating only a subset of experts per input. However, we observe that the commonly used auxiliary load balancing loss often leads to expert overlap and overly uniform routing, which hinders expert specialization and degrades overall performance during post-training. To address this, we propose a simple yet effective solution that introduces two complementary objectives: (1) an orthogonality loss to enco...

Relationship Analysis

Both papers belong to the Auxiliary Loss-Based Alignment category, using supplementary loss functions to improve router-expert alignment in MoE models. They overlap in addressing the misalignment between router decisions and expert capabilities through auxiliary training objectives. However, the original paper introduces a coupling loss based on perturbed router embeddings as proxy tokens and intermediate activation norms, while the candidate paper focuses on orthogonality loss to reduce expert overlap and variance loss to diversify routing decisions, targeting the expert homogenization problem caused by load balancing.

2. On the representation collapse of sparse mixture of experts

Authors: Chi, Zewen, Dong Li, Zewen Chi, Huang, et al. (31 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Sparse mixture of experts provides larger model capacity while requiring a constant computational overhead. It employs the routing mechanism to distribute input tokens to the best-matched experts according to their hidden representations. However, learning such a routing mechanism encourages token clustering around expert centroids, implying a trend toward representation collapse. In this work, we propose to estimate the routing scores between tokens and experts on a low-dimensional hypersphere....

Relationship Analysis

Both papers belong to the Auxiliary Loss-Based Alignment category, using supplementary loss functions to improve router-expert alignment in MoE models. The original paper's ERC loss enforces alignment by measuring intermediate activation norms between perturbed router embeddings and experts, while the candidate paper addresses representation collapse through dimensionality reduction and L2 normalization of routing scores on a hypersphere. The key difference is that the original paper focuses on coupling through activation-based constraints with proxy tokens, whereas the candidate paper tackles the geometric properties of the routing space to prevent collapse.

Contributions Analysis

Overall novelty summary. The paper proposes an expert-router coupling loss (ERC loss) that enforces bidirectional alignment between routing decisions and expert capabilities through proxy tokens and contrastive constraints. It resides in the Auxiliary Loss-Based Alignment leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader Router-Expert Alignment Mechanisms branch. This positioning suggests the work addresses a recognized but not heavily explored approach—using supplementary loss functions to couple routers and experts—rather than entering a crowded subfield.

The taxonomy reveals neighboring approaches in sibling leaves: Architectural Coupling Mechanisms (two papers) integrates structural constraints rather than auxiliary losses, while Representation and Manifold Alignment (two papers) focuses on aligning routing weight manifolds with task embeddings. The broader Router Design and Optimization Strategies branch contains more populous leaves like Dynamic and Adaptive Routing (four papers) and Domain-Specific Routing (five papers), which pursue routing improvements through architectural innovation rather than explicit alignment objectives. The ERC loss approach thus occupies a distinct methodological niche, emphasizing training-time loss functions over architectural redesign or domain specialization.

Among twenty candidates examined, neither contribution shows clear refutation. The ERC loss mechanism (ten candidates examined, zero refutable) and its use for studying expert specialization (ten candidates examined, zero refutable) both appear to introduce novel formulations within the limited search scope. The bidirectional constraint design—requiring both experts to prefer their proxy tokens and proxy tokens to prefer their designated experts—does not appear directly anticipated in the examined prior work, though the small candidate pool and sparse taxonomy leaf suggest this assessment reflects top-twenty semantic matches rather than exhaustive coverage.

Based on the limited literature search and sparse taxonomy positioning, the work appears to contribute a distinct auxiliary loss formulation to a relatively underexplored alignment strategy. The analysis covers top-twenty semantic candidates and does not claim exhaustive field coverage, particularly given the small number of papers in the Auxiliary Loss-Based Alignment leaf and adjacent leaves. The novelty assessment reflects this bounded search scope rather than a comprehensive survey of all MoE alignment techniques.

This paper presents **2 main contributions**, each analyzed against relevant prior work:

Contribution 1: Expert-router coupling loss (ERC loss)

Description: The authors introduce a lightweight auxiliary loss that couples expert capabilities with router decisions by treating router parameters as cluster centers, perturbing them to create proxy tokens, and enforcing constraints that ensure each expert is most activated by its designated proxy token and vice versa. This optimization strengthens the alignment between routing decisions and expert capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MoE at Scale: From Modular Design to Deployment in Large-Scale Machine Learning Systems

URL: View paper

Brief Assessment

MoE at Scale[13] provides a survey categorizing MoE architectures along dimensions like gating mechanisms and expert sparsity, but does not propose auxiliary losses for coupling routers and experts or methods similar to the ERC loss described in the original paper.

2. Ta-moe: Topology-aware large scale mixture-of-expert training

URL: View paper

Brief Assessment

Topology Aware Training[60] focuses on topology-aware routing to optimize communication patterns in distributed MoE training, not on coupling expert capabilities with router decisions through auxiliary losses.

3. Enhancing the" Immunity" of Mixture-of-Experts Networks for Adversarial Defense

URL: View paper

Brief Assessment

Immunity Adversarial Defense[59] focuses on adversarial robustness through mutual information and position stability losses applied to expert heatmaps in a mixture-of-experts framework for defense against attacks. It does not address the router-expert coupling problem or propose auxiliary losses that align routing decisions with expert capabilities as in the original paper.

4. Uncertainty prediction and calibration using multi-expert gating mechanism

URL: View paper

Brief Assessment

Multi Expert Gating[62] focuses on uncertainty prediction and calibration in time series models using multi-expert gating mechanisms, not on coupling routers and experts in mixture-of-experts language models through auxiliary losses.

5. A survey on mixture of experts: Advancements, challenges, and future directions

URL: View paper

Brief Assessment

Advancements Challenges Directions[57] is a survey paper that provides an overview of MoE architectures. It does not propose novel auxiliary losses for coupling routers and experts.

6. Enhancing molecular property prediction via mixture of collaborative experts

URL: View paper

Brief Assessment

Molecular Property Prediction[61] focuses on molecular property prediction using mixture of experts for graph classification tasks, not on general mixture-of-experts language models. The candidate's expert-specific loss addresses decision dominance within expert groups in molecular prediction contexts, which is technically distinct from the original paper's ERC loss that couples router parameters as cluster centers with expert capabilities through proxy tokens and intermediate activation norms in MoE-LLMs.

7. Leave It to the Experts: Detecting Knowledge Distillation via MoE Expert Signatures

URL: View paper

Brief Assessment

Expert Signatures Detection[65] focuses on detecting knowledge distillation by analyzing MoE routing patterns as fingerprints, not on designing auxiliary losses to improve expert-router coupling during training. The candidate's routing analysis serves a completely different purpose (distillation detection) than the original's training optimization objective.

8. Learning in gated neural networks

URL: View paper

Brief Assessment

Gated Neural Networks[64] focuses on learning parameters in mixture-of-experts models through specially designed loss functions (L4 for regressors, Llog for gating parameters) but does not address the specific problem of coupling expert capabilities with router decisions through auxiliary losses as proposed in the original paper. The candidate's approach learns regressors and gating parameters separately using different loss functions, while the original paper's ERC loss explicitly couples these components through proxy tokens and activation norms.

9. Beyond Degradation Conditions: All-in-One Image Restoration via HOG Transformers

URL: View paper

Brief Assessment

HOG Transformers Restoration[58] focuses on image restoration using histogram of oriented gradients (HOG) features with a HOG loss for structural fidelity, not on mixture-of-experts models or router-expert coupling mechanisms in language models.

10. Distributionally-Robust Gradient Routing: A Bilevel Sparse Optimization Problem for Compute-Aware Mixture-of-Experts Training

URL: View paper

Brief Assessment

Distributionally Robust Routing[63] focuses on bilevel optimization for compute-aware routing under domain uncertainty using f-divergence ambiguity sets and gradient-traffic regularization. It does not propose an auxiliary loss that couples expert capabilities with router decisions through proxy tokens and perturbations as described in the ERC loss.

Contribution 2: ERC loss as a tool for studying expert specialization

Description: The ERC loss enables flexible control and quantitative tracking of expert specialization levels during training through the hyperparameter alpha and the noise bound epsilon. This capability allows researchers to investigate the trade-off between specialization and model performance, challenging previous beliefs about expert orthogonality derived from small-scale experiments.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Theory of mixture-of-experts for mobile edge computing

URL: View paper

Brief Assessment

Mobile Edge Computing[56] focuses on expert specialization in mobile edge computing networks for continual learning tasks, not on studying expert specialization mechanisms in general MoE models through auxiliary loss design.

2. Advancing Expert Specialization for Better MoE

URL: View paper

Brief Assessment

Advancing Expert Specialization[7] focuses on orthogonality and variance losses to enhance expert specialization during post-training/fine-tuning, while the original paper's ERC loss provides a controllable framework (via alpha and epsilon) for investigating specialization-performance trade-offs during pre-training. The candidate does not demonstrate prior work on using auxiliary losses as quantitative research tools for studying expert specialization dynamics.

3. Omoe: Diversifying mixture of low-rank adaptation by orthogonal finetuning

URL: View paper

Brief Assessment

Orthogonal Finetuning Diversifying[52] focuses on orthogonalizing expert representations in LoRA-based MoE for parameter-efficient fine-tuning, not on developing tools to study expert specialization dynamics during pre-training with quantitative control mechanisms like the ERC loss's alpha and epsilon parameters.

4. Superposition in Mixture of Experts

URL: View paper

Brief Assessment

Superposition MoE[51] focuses on measuring superposition and monosemanticity in toy MoE models through feature dimensionality metrics, not on controlling or tracking expert specialization during training via auxiliary losses like the ERC loss.

5. Plant disease classification in the wild using vision transformers and mixture of experts

URL: View paper

Brief Assessment

Plant Disease Vision[53] focuses on applying mixture of experts to plant disease classification in agricultural settings, not on developing tools for studying expert specialization dynamics or the specialization-performance trade-off in MoE models.

6. Vimoe: An empirical study of designing vision mixture-of-experts

URL: View paper

Brief Assessment

Vision MoE Design[34] focuses on architectural design choices for vision transformers with MoE layers, using shared experts for stability. It does not propose a loss function like ERC that enables flexible control of specialization through hyperparameters or quantitative tracking via noise bounds.

7. A survey on mixture of experts: Advancements, challenges, and future directions

URL: View paper

Brief Assessment

Advancements Challenges Directions[57] surveys existing MoE methods but does not introduce tools for quantitatively controlling or tracking expert specialization during training.

8. Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts

URL: View paper

Brief Assessment

Diversifying Expert Knowledge[54] focuses on pruning redundant experts in MoE models by grouping similar experts based on their knowledge overlap, rather than studying expert specialization dynamics during training or investigating the trade-off between specialization and performance through controllable hyperparameters.

9. A closer look into mixture-of-experts in large language models

URL: View paper

Brief Assessment

Closer Look MoE[4] focuses on analyzing existing MoE models through parameter and behavioral similarity measurements, rather than proposing a controllable mechanism for studying specialization during training. The paper does not present a loss function or training methodology that enables flexible control of expert specialization levels.

10. Unifying mixture of experts and multi-head latent attention for efficient language models

URL: View paper

Brief Assessment

Multi Head Latent[55] focuses on combining MoE with multi-head latent attention for memory efficiency in small language models, not on studying expert specialization dynamics or developing tools to quantitatively track specialization levels during training.

Appendix: Text Similarity Detection

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

References

- [0] Coupling Experts and Routers in Mixture-of-Experts via an Auxiliary Loss View paper
- [1] On the benefits of learning to route in mixture-of-experts models View paper
- [2] Mixture-of-experts with expert choice routing View paper
- [3] A survey on inference optimization techniques for mixture of experts models View paper
- [4] A closer look into mixture-of-experts in large language models View paper
- [5] Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing View paper
- [6] Novel token-level recurrent routing for enhanced mixture-of-experts performance View paper
- [7] Advancing Expert Specialization for Better MoE View paper
- [8] Learning to route among specialized experts for zero-shot generalization View paper

- [9] Upcycling Instruction Tuning from Dense to Mixture-of-Experts via Parameter Merging View paper
- [10] Routing experts: Learning to route dynamic experts in multi-modal large language models View paper
- [11] Decoding Knowledge Attribution in Mixture-of-Experts: A Framework of Basic-Refinement Collaboration and Efficiency Analysis View paper
- [12] ROUTERRETRIEVER: Routing over a Mixture of Expert Embedding Models View paper
- [13] MoE at Scale: From Modular Design to Deployment in Large-Scale Machine Learning Systems View paper
- [14] Training mixture-of-experts: A focus on expert-token matching View paper
- [15] Mixture of Routers View paper
- [16] Enabling MoE on the Edge via Importance-Driven Expert Scheduling View paper
- [17] BadMoE: Backdooring Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts View paper
- [18] Frequency-Augmented Mixture-of-Heterogeneous-Experts Framework for Sequential Recommendation View paper
- [19] Not All Models Suit Expert Offloading: On Local Routing Consistency of Mixture-of-Expert Models View paper
- [20] Symbolic Mixture-of-Experts: Adaptive Skill-based Routing for Heterogeneous Reasoning View paper
- [21] ExpertFlow: Adaptive Expert Scheduling and Memory Coordination for Efficient MoE Inference View paper
- [22] Rewiring Experts on the Fly: Continuous Rerouting for Better Online Adaptation in Mixture-of-Expert models View paper
- [23] Chain-of-Experts: Unlocking the Communication Power of Mixture-of-Experts Models View paper
- [24] From Score Distributions to Balance: Plug-and-Play Mixture-of-Experts Routing View paper
- [25] On the representation collapse of sparse mixture of experts View paper
- [26] Q-moe: Connector for mllms with text-driven routing View paper
- [27] Multilingual Routing in Mixture-of-Experts View paper
- [28] DA-MoE: Towards Dynamic Expert Allocation for Mixture-of-Experts Models View paper
- [29] The evolution of moe: A survey from basics to breakthroughs View paper
- [30] Routing Manifold Alignment Improves Generalization of Mixture-of-Experts LLMs View paper
- [31] Part-Of-Speech Sensitivity of Routers in Mixture of Experts Models View paper
- [32] Mixture of experts for network optimization: a large language model-enabled approach View paper
- [33] The evolution of mixture of experts: A survey from basics to breakthroughs View paper
- [34] Vimoe: An empirical study of designing vision mixture-of-experts View paper
- [35] GM-MoE: Low-Light Enhancement with Gated-Mechanism Mixture-of-Experts View paper
- [36] : Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design View paper
- [37] Sparse moe with language guided routing for multilingual machine translation View paper
- [38] Router Upcycling: Leveraging Mixture-of-Routers in Mixture-of-Experts Upcycling View paper
- [39] OrdMoE: Preference Alignment via Hierarchical Expert Group Ranking in Multimodal Mixture-of-Experts LLMs View paper
- [40] AT-MoE: Adaptive Task-planning Mixture of Experts via LoRA Approach View paper
- [41] Steer-MoE: Efficient Audio-Language Alignment with a Mixture-of-Experts Steering Module View paper
- [42] How Do Shared Experts Dynamically Adapt to Routing Constraints in Mixture-of-Experts? View paper
- [43] Layerwise Recurrent Router for Mixture-of-Experts View paper
- [44] DRAMoE: Boosting Adversarial Robustness with Adversarial Training and Adaptive Mixture of Experts View paper
- [45] Tight Clusters Make Specialized Experts View paper
- [46] Efficient Routing in Sparse Mixture-of-Experts View paper
- [47] Improving Expert Specialization in Mixture of Experts View paper
- [48] Towards Using Partitioned GPU Virtual Functions for Mixture of Experts View paper
- [49] What really matters for person re-identification? A Mixture-of-Experts Framework for Semantic Attribute Importance View paper
- [50] ExpertFlow: Optimized Expert Activation and Token Allocation for Efficient Mixture-of-Experts Inference View paper
- [51] Superposition in Mixture of Experts View paper
- [52] Omoe: Diversifying mixture of low-rank adaptation by orthogonal finetuning View paper
- [53] Plant disease classification in the wild using vision transformers and mixture of experts View paper
- [54] Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts View paper
- [55] Unifying mixture of experts and multi-head latent attention for efficient language models View paper
- [56] Theory of mixture-of-experts for mobile edge computing View paper
- [57] A survey on mixture of experts: Advancements, challenges, and future directions View paper
- [58] Beyond Degradation Conditions: All-in-One Image Restoration via HOG Transformers View paper
- [59] Enhancing the" Immunity" of Mixture-of-Experts Networks for Adversarial Defense View paper
- [60] Ta-moe: Topology-aware large scale mixture-of-expert training View paper
- [61] Enhancing molecular property prediction via mixture of collaborative experts View paper
- [62] Uncertainty prediction and calibration using multi-expert gating mechanism View paper
- [63] Distributionally-Robust Gradient Routing: A Bilevel Sparse Optimization Problem for Compute-Aware Mixture-of-Experts Training View paper
- [64] Learning in gated neural networks View paper
- [65] Leave It to the Experts: Detecting Knowledge Distillation via MoE Expert Signatures View paper