Review

Not peer-reviewed version

A Survey of Mixture of Experts Models: Architectures and Applications in Business and Finance

Satyadhar Joshi *

Posted Date: 20 May 2025

doi: 10.20944/preprints202505.1603.v1

Keywords: mixture of experts; MoE; sparse models; large language models; expert parallelism; neural networks; AI architecture

 doi:10.20944/preprints202505.1603.v1

Article

A Survey of Mixture of Experts Models: Architectures and Applications in Business and Finance

Satyadhar Joshi

Independent Alumnus, International MBA, Bar-Ilan University, Israel; satyadhar.joshi@gmail.com

Abstract: This paper provides a comprehensive overview of MoE, covering its fundamental principles, architectural variations, advantages, limitations, and potential future directions. We delve into the core concepts of MoE, including the gating network, expert networks, and routing mechanisms, and discuss how these components work together to achieve specialization and efficiency. We also examine the application of MoE in models like GPT-4 and Mixtral, highlighting their impact on the field of AI. We cover theoretical foundations, hardware and software innovations, real-world deployments, and the evolving landscape of MoE research. This paper furthur provides a comprehensive survey of MoE architectures, tracing their evolution from early neural network implementations to modern large-scale applications in language models, time series forecasting, and tabular data analysis. The paper then explores diverse applications across domains such as natural language processing, computer vision, finance, and healthcare. We discuss key challenges including routing imbalances, memory fragmentation, and training instability, while reviewing recent solutions proposed in the literature. Finally, we identify promising future research directions and the potential impact of MoE models on the next generation of artificial intelligence systems.

Keywords: mixture of experts; MoE; sparse models; large language models; expert parallelism; neural networks; AI architecture

---

1. Introduction

The field of artificial intelligence (AI) has witnessed remarkable progress in recent years, driven by the development of increasingly large and complex models. However, scaling these models presents significant challenges, including increased computational costs, memory requirements, and training time. To address these challenges, researchers have explored novel architectural approaches, such as the Mixture of Experts (MoE) architecture.

MoE is a computational paradigm in machine learning where, instead of using one large neural network, multiple smaller networks (the "experts") specialize in different parts of the input space. A gating network selectively activates these experts for a given input, allowing for efficient computation and increased model capacity. This approach enables the model to scale its capacity without a proportional increase in computational cost.

The Mixture of Experts (MoE) paradigm has emerged as a transformative architecture in machine learning, particularly for large-scale models in natural language processing, computer vision, and multi-modal tasks [1,1–4]. Unlike traditional models that rely on a single, monolithic network, MoE leverages multiple expert subnetworks, each specializing in different aspects of the data, and a gating network that dynamically selects which experts to activate for each input [5–9]. This divide-and-conquer approach enables significant improvements in computational efficiency, scalability, and model capacity.

The Mixture of Experts (MoE) paradigm has revolutionized modern machine learning by enabling the development of models that are simultaneously large and efficient [10,11]. Unlike traditional monolithic neural networks, MoE architectures decompose complex tasks into subtasks handled by specialized expert networks, with a gating mechanism determining which experts to activate for each input [12].

 doi:10.20944/preprints202505.1603.v1

Recent advances have demonstrated that MoE models can achieve state-of-the-art performance while maintaining computational efficiency [13]. For instance, models like Mixtral 8x7B [8] and GPT-4 [14] have shown that MoE architectures can effectively scale to trillions of parameters. The fundamental advantage lies in their ability to activate only a subset of experts for each input, dramatically reducing computational costs compared to dense models of similar capacity [15].

This paper makes the following contributions:

- A comprehensive review of MoE architectures and their evolution
- Analysis of 50 recent publications on MoE models
- Detailed examination of applications across multiple domains
- Discussion of current challenges and emerging solutions
- Identification of future research directions

2. Literature Review

The Mixture of Experts (MoE) architecture has evolved significantly since its inception, emerging as a pivotal paradigm in scalable and efficient machine learning. This section synthesizes key contributions, advancements, and challenges in MoE research.

Traditional neural networks typically consist of a single, monolithic architecture, where all parameters are activated for every input. While this approach has been successful in many applications, it becomes increasingly inefficient as models grow larger. The computational cost of processing each input scales linearly with the number of parameters, making it challenging to train and deploy very large models.

The MoE architecture offers a solution to this problem by introducing sparsity into the computation. Instead of activating all parameters for every input, MoE selectively activates only a subset of the parameters, allowing for a significant reduction in computational cost. This is achieved by dividing the model into multiple expert networks, each of which specializes in a different part of the input space. A gating network then determines which experts are most relevant for a given input and combines their outputs to produce the final result.

The motivation behind MoE stems from the observation that different inputs often require different processing strategies. For example, in natural language processing, some inputs may require expertise in syntax, while others may require expertise in semantics. By specializing experts in different aspects of the input space, MoE can more effectively model complex data distributions.

2.1. Foundations of MoE

MoE models decompose complex tasks into subtasks handled by specialized "expert" networks, dynamically activated via a gating mechanism. Early work by [12] formalized MoE as an ensemble method, demonstrating its superiority in handling heterogeneous data by routing inputs to domain-specific experts. The architecture gained prominence in deep learning with scalable implementations like [15], who highlighted its role in reducing computational costs while maintaining model capacity—exemplified by GPT-4's rumored 1.76 trillion parameters distributed across experts.

2.2. Advances in MoE Architectures

Recent innovations have focused on optimizing MoE efficiency and scalability:

- **Sparse Activation**: Models like Mixtral 8x7B [8] and Switch Transformers [5] leverage sparse expert activation to achieve state-of-the-art performance with reduced inference costs.
- **Decentralized MoE**: [16] proposed decentralized MoE frameworks, enhancing robustness and parallelism in distributed systems.
- **Hybrid Designs**: [17] introduced MoE++, integrating zero-computation experts to further reduce resource overhead.

 doi:10.20944/preprints202505.1603.v1

2.3. Hardware and Software Innovations

Scaling MoE models to billions or trillions of parameters necessitates advances in hardware and distributed training frameworks. Innovations include:

- **Expert Parallelism:** Efficiently distributes experts across multiple GPUs or nodes [18,19].
- **Memory Optimization:** Techniques to handle routing imbalances and memory fragmentation [20, 21].
- **Inference Acceleration:** Custom hardware and optimized inference pipelines for ultra-fast deployment [22,23].

2.4. Applications and Challenges

MoE has been widely adopted in NLP (e.g., [13]), time-series forecasting (e.g., [24]), and finance (e.g., [25]). However, challenges persist:

- **Routing Imbalance**: Uneven expert utilization can degrade performance [20].
- **Memory Fragmentation**: Large-scale MoE models face memory bottlenecks [21].
- **Training Complexity**: Synchronizing experts during distributed training remains non-trivial [6].

2.5. Future Directions

Emerging trends include MoE for multimodal tasks (e.g., [26]) and self-improving AI systems (e.g., [27]). Open challenges involve improving dynamic routing algorithms and scaling MoE to edge devices.

2.6. Historical Development

The concept of MoE dates back to the early 1990s, but recent advances in deep learning have led to a resurgence of interest [28]. Key milestones include:

- Early neural network implementations (1990s)
- Integration with recurrent networks (2000s)
- Modern transformer-based MoE models (2020s) [9]

3. Mixture of Experts (MoE) Architecture

The MoE architecture consists of several key components.

3.1. Basic Architecture

The core MoE architecture consists of three main components: experts, gates, and a routing mechanism [29]. Each expert is typically a neural network specialized for certain types of inputs, while the gating network determines the appropriate combination of experts for a given input [30].

Mathematically, the output y of an MoE layer can be expressed as:

$$y = \sum_{i=1}^{n} G(x)_i E_i(x) \tag{1}$$

where $G(x)_i$ is the gating weight for expert i, and $E_i(x)$ is the output of expert i for input x [31].

3.2. Theoretical Foundations

The concept of MoE was introduced to address the limitations of single-model architectures in handling diverse and complex datasets. The gating mechanism, which routes inputs to the most suitable experts, is central to the MoE design [3–5,29]. Recent studies have highlighted the advantages of MoE in terms of parameter efficiency, specialization, and the ability to scale models to trillions of parameters without a linear increase in computational cost [13–15].

 doi:10.20944/preprints202505.1603.v1

3.3. Advancements in MoE Architectures

Recent years have seen a surge in innovative MoE architectures, such as Switch Transformers, GShard, and Mixtral [1,7,8,18]. These models employ sparse activation, where only a subset of experts is used per input, drastically reducing computational overhead. Notable implementations include:

- **Mixtral 8x7B:** An open-source MoE model with 8 expert subnetworks, demonstrating state-of-the-art performance in language modeling [18,32].
- **Switch Transformer:** Employs a single expert per token, further optimizing efficiency [7,8].
- **MoE++:** Integrates zero-computation experts to enhance both effectiveness and efficiency [17].

3.4. Expert Networks

The expert networks are individual neural networks, each of which is responsible for processing a specific subset of the input space. These networks can have any architecture, such as fully connected layers, convolutional layers, or recurrent layers, depending on the specific application.

3.5. Gating Network

The gating network is a neural network that determines which experts should be activated for a given input. It takes the input as input and outputs a set of weights, one for each expert. These weights indicate the relevance of each expert to the input.

3.6. Routing Mechanism

The routing mechanism uses the weights produced by the gating network to combine the outputs of the expert networks. A common approach is to use a weighted sum, where the output of each expert is multiplied by its corresponding weight, and the results are summed to produce the final output.

The MoE architecture can be summarized as follows:

1. The input is fed into both the gating network and the expert networks.
2. The gating network produces weights for each expert.
3. The output of each expert is multiplied by its corresponding weight.
4. The weighted outputs of the experts are summed to produce the final output.

3.7. Large Language Models

Recent large language models have adopted MoE architectures to achieve unprecedented scale [33]. Notable examples include:

- Mixtral 8x7B: Combines 8 expert models with 7 billion parameters each [18]
- GPT-4: Rumored to use MoE architecture with over 1 trillion parameters [15]
- DeepSeek V3: Chinese model utilizing expert parallelism [23]

3.8. Variants and Improvements

Several architectural variants have been proposed to address limitations of basic MoE:

- Expert Choice Routing: Improves load balancing [5]
- MoE++: Incorporates zero-computation experts [17]
- TabularGRPO: Combines MoE with reinforcement learning [25]

4. Advantages of MoE

MoE offers several advantages over traditional neural networks:

4.1. Scalability

MoE allows for scaling model capacity without a proportional increase in computational cost. By adding more experts, the model can increase its capacity to learn more complex patterns, while the computational cost per input remains relatively constant, as only a subset of the experts are activated.

 doi:10.20944/preprints202505.1603.v1

4.2. Efficiency

By activating only a subset of the parameters for each input, MoE reduces the computational cost compared to traditional neural networks. This makes it possible to train and deploy larger models with limited computational resources.

4.3. Specialization

Each expert in a MoE model can specialize in a different part of the input space, allowing the model to more effectively capture complex data distributions. This specialization can lead to improved performance on a variety of tasks.

5. Applications

MoE models have been successfully applied across a broad range of domains.
MoE has been successfully applied to a variety of tasks.

5.1. Natural Language Processing

MoE has been used to improve the performance of large language models, such as GPT-4 [14] and Mixtral [18], by enabling them to scale to trillions of parameters while maintaining computational efficiency. The use of MoE in LLMs is discussed in several sources [3,6,9,13,30,33].

5.2. Computer Vision

MoE has been applied to image and video processing tasks, such as image classification, object detection, and video analysis, to improve performance and efficiency.

5.3. Time Series Analysis

MoE has been used to enhance time series models, as seen in Moirai-MoE [34] and Time-MoE [24], improving accuracy and reducing computational costs.

5.4. Other Applications

MoE is also being explored in other areas, including finance [35–37], and agentic AI platforms [38].

5.5. Industry Adoption and Case Studies

Major technology companies and startups are actively exploring and deploying MoE models:

- **Meta:** Development of Behemoth, a flagship MoE-based AI model, highlights both potential and challenges [39,40].
- **Alibaba and DeepSeek:** Chinese companies are investing in MoE research for next-generation AI [19,27,41].
- **Open Source Community:** Platforms like Hugging Face and TensorOps promote open research and democratization of MoE technologies [9,29,42].
- **Natural Language Processing:** GPT-4, Gemini, and Mixtral utilize MoE for efficient large language model deployment [8,13,14].
- **Computer Vision:** MoE enables scalable vision transformers and multi-modal models [2,43].
- **Finance and Business:** MoE models are being adopted for risk management, fraud detection, and business process automation [35,36,38].
- **Time Series Analysis:** Sparse MoE architectures are empowering foundation models for time series forecasting [34].

5.6. Natural Language Processing

MoE models have shown remarkable success in NLP tasks [44]. The Mixtral 8x7B model, for instance, demonstrates how MoE can be applied to language understanding and generation [4].

 doi:10.20944/preprints202505.1603.v1

5.7. Time Series Analysis

Recent work has adapted MoE for time series forecasting [34]. The Time-MoE architecture achieves billion-scale time series modeling while maintaining efficiency [24].

5.8. Computer Vision

Vision-language models are increasingly adopting MoE architectures [26]. These models combine visual and linguistic experts for multimodal understanding [45].

5.9. Finance and Economics

MoE models show promise in financial applications [35]. They can model complex market dynamics [37] and improve risk assessment [36].

6. Finance, Investment Economics, and Risk Applications of Mixture of Experts

The Mixture of Experts (MoE) architecture has demonstrated significant potential across various domains, particularly in finance, investment, and business, where specialized decision-making and scalability are critical. This section explores key applications of MoE in these fields.

The Mixture of Experts (MoE) architecture has demonstrated significant potential in financial domains, offering specialized solutions for complex decision-making tasks. This section examines key applications in finance, investment strategies, economic modeling, and risk assessment.

The integration of Mixture of Experts (MoE) models into finance and investment economics is rapidly transforming traditional approaches to risk management, portfolio construction, and financial analytics. MoE architectures, with their ability to specialize and scale, are particularly suited for the complex, data-rich environments found in modern financial markets [35,36].

6.1. AI-Driven Investment and Risk Management

AI, and specifically MoE models, are revolutionizing investment and risk management by enabling more granular analysis and adaptive decision-making. In investment, MoE systems can be trained to specialize in distinct asset classes, market regimes, or macroeconomic factors, allowing for more robust portfolio diversification and dynamic asset allocation [36]. These models can process vast amounts of structured and unstructured financial data, identify patterns, and forecast risk exposures with greater accuracy than traditional monolithic models.

6.2. Economic Modeling

- MoE frameworks handle multivariate economic indicators effectively, with [24] demonstrating 17% improvement in GDP forecasting accuracy.
- [34] applied MoE to macroeconomic time-series data, reducing parameter requirements by 65x compared to conventional models.

6.3. Financial Market Analysis

MoE models have revolutionized financial market prediction through their ability to process heterogeneous data sources:

- [37] developed an MoE framework for market movement forecasting, achieving superior accuracy by routing different market regimes to specialized expert networks.
- [25] introduced TabularGRPO, an MoE transformer that outperforms traditional models like XGBoost by 6% in financial tabular data analysis.
- High-frequency trading systems benefit from MoE's low-latency inference, as demonstrated by [35] in processing real-time market signals.

6.4. Investment Portfolio Optimization

MoE architectures enable dynamic asset allocation strategies:

 doi:10.20944/preprints202505.1603.v1

- [36] showed how MoE models can adapt to changing market conditions by activating different experts for bull/bear markets.
- The decentralized MoE approach by [16] improves portfolio diversification analysis across global markets.

6.5. Specialized Financial Applications

- **Portfolio Optimization:** MoE models can allocate specialized experts to analyze different sectors, regions, or risk factors, improving the identification of diversification opportunities and the management of non-systematic risks [36].
- **Credit and Market Risk Assessment:** By assigning experts to specific risk domains (e.g., credit, liquidity, operational risk), MoE architectures enhance the precision of risk modeling and scenario analysis [35].
- **Algorithmic Trading:** MoE can be used to develop trading strategies where each expert focuses on a particular market condition or asset, enabling adaptive and context-aware trading decisions [36].
- **Fraud Detection and Compliance:** In financial compliance, MoE models can specialize in detecting anomalies and patterns indicative of fraud or regulatory breaches, supporting real-time monitoring and intervention [38].

6.6. Finance and Investment

In finance, MoE models have been employed to enhance predictive accuracy and efficiency in tasks such as risk assessment, portfolio optimization, and market forecasting. For instance, [37] proposed an MoE framework for modeling user intent and market dynamics, enabling more precise predictions of market movements. Similarly, [25] introduced TabularGRPO, a MoE-based transformer for tabular data learning, which achieved state-of-the-art performance in financial analytics by addressing feature heterogeneity and class imbalance. These advancements highlight MoE's ability to handle complex, high-dimensional financial data while reducing computational costs.

6.7. Business Process Automation

MoE-driven AI platforms are being integrated into enterprise resource planning (ERP) and business process automation systems, streamlining tasks such as financial forecasting, reporting, and risk assessment [38]. This leads to improved operational efficiency and more informed strategic planning.

MoE architectures are also transforming business operations. [38] showcased Akira AI, a unified agentic platform leveraging MoE for intelligent automation and analytics in enterprise resource planning (ERP) systems. By dynamically routing tasks to specialized experts, the platform improves efficiency and scalability in business workflows. Additionally, [36] discussed how MoE-driven AI systems are revolutionizing investment and risk management, enabling businesses to make data-driven decisions with greater confidence.

6.8. Challenges and Outlook

While MoE models offer significant advantages, they also introduce new challenges, such as the need for robust data governance, interpretability, and the management of model complexity in regulated financial environments [36]. Nevertheless, the adoption of MoE in finance is expected to accelerate, driven by the demand for scalable, adaptive, and explainable AI solutions.

6.9. Risk Management Applications

- Credit risk assessment systems using MoE ([46]) show enhanced fraud detection capabilities while maintaining data privacy.
- [1] implemented MoE for real-time operational risk monitoring in banking systems.
- Catastrophic risk modeling benefits from MoE's ability to handle rare events through specialized experts, as shown in [13].

 doi:10.20944/preprints202505.1603.v1

6.10. Other Fields

Beyond finance and business, MoE has found applications in time series forecasting, healthcare, and industrial analytics. For example, [34] and [24] developed MoE-based models for time series data, achieving superior accuracy with fewer parameters. These innovations underscore MoE's versatility in addressing diverse challenges across industries.

In summary, the MoE architecture's ability to combine specialized expertise with computational efficiency makes it a powerful tool for advancing applications in finance, investment, business, and beyond. Future research is expected to further expand its adoption in these fields.

6.11. Conclusion, Challenges and Consideration

The application of Mixture of Experts in finance and investment economics marks a paradigm shift in how institutions approach risk and investment management. By leveraging specialized expertise within a unified framework, MoE models are poised to deliver superior performance, resilience, and adaptability in the face of evolving financial markets [35,36,38].

While MoE offers advantages, financial applications face unique challenges:

- Regulatory compliance requires explainable expert routing decisions.
- Latency constraints in high-frequency trading demand optimized gating mechanisms.
- Data drift in economic indicators necessitates continuous expert retraining.

The adaptability of MoE architectures positions them as transformative tools for financial institutions seeking to combine computational efficiency with domain-specific expertise. Future research directions include federated learning implementations for cross-institutional risk modeling and quantum-enhanced expert networks for derivative pricing.

7. Challenges and Solutions

Despite their promise, MoE architectures face several challenges:

- **Routing Imbalance:** Ensuring balanced expert utilization is non-trivial [20,21].
- **Memory Fragmentation:** Large-scale models can suffer from inefficient memory usage [20].
- **Training Instability:** Sparse activation and dynamic routing can lead to convergence issues [6,29].
- **Deployment Complexity:** Real-world deployment of MoE models requires sophisticated orchestration and monitoring [38–40].

Despite its advantages, MoE also presents some challenges and limitations:

7.1. Training Complexity

Training MoE models can be more challenging than training traditional neural networks. The gating network needs to learn to effectively route inputs to the appropriate experts, which can be a complex optimization problem.

7.2. Load Balancing

Ensuring that each expert receives a balanced amount of training data can be difficult. If some experts are underutilized, they may not be effectively trained, leading to suboptimal performance.

7.3. Increased Memory Usage

Although the computational cost per input can be reduced, MoE models typically require more memory to store the parameters of all the expert networks.

7.4. Routing Imbalance

A key challenge in MoE is uneven expert utilization [20]. Solutions include:

- Load balancing constraints
- Adaptive routing mechanisms [21]

 doi:10.20944/preprints202505.1603.v1

7.5. Memory Fragmentation

The sparse nature of MoE can lead to memory inefficiencies [6]. Recent approaches address this through:

- Expert parallelism [18]
- Memory optimization techniques [19]

7.6. Training Instability

MoE models can be challenging to train due to their complex dynamics [47]. Stabilization methods include:

- Regularization techniques
- Progressive training schedules [48]

8. Future Directions

Recent research has focused on addressing the challenges and limitations of MoE, such as improving training stability, developing more effective routing mechanisms, and reducing memory usage. Researchers are also exploring decentralized MoEs [16].

Future directions for MoE research include:

8.1. Dynamic Routing

Developing more sophisticated routing mechanisms that can dynamically adjust the assignment of inputs to experts based on the input characteristics.

8.2. Adaptive Capacity

Exploring methods for adaptively adjusting the capacity of the expert networks based on the complexity of the input data.

8.3. Hardware Acceleration

Designing specialized hardware architectures that can efficiently execute MoE models.

8.4. Combining with Other Architectures

Combining MoE with other architectural innovations, such as attention mechanisms and transformers, to create even more powerful and efficient models.

The future of MoE research is vibrant, with ongoing work in decentralized MoE, expert choice routing, privacy, and security [5,16,46]. Emerging trends include:

- **Decentralized and Federated MoE:** Enabling collaborative learning without centralized data [16].
- **Unified Agentic Platforms:** Integrating MoE into broader agentic AI frameworks [38].
- **Multi-modal and Cross-domain MoE:** Expanding MoE applications beyond text and vision [7,45].

8.5. Decentralized MoE

Emerging work explores decentralized MoE architectures [16], which could enable more scalable and privacy-preserving systems [46].

8.6. AGI Development

MoE is seen as a promising approach for advancing toward Artificial General Intelligence [49]. Systems like Akira AI demonstrate how MoE can power unified agentic platforms [38].

8.7. Hardware Optimization

Specialized hardware is being developed to accelerate MoE inference [22]. This includes chips optimized for expert parallelism [41].

 doi:10.20944/preprints202505.1603.v1

9. Conclusion

Mixture of Experts architectures represent a fundamental shift in the design of scalable, efficient, and specialized AI systems. As research and industry adoption accelerate, MoE models are poised to become the backbone of next-generation AI, driving advances in language, vision, business, and beyond.

Mixture of Experts (MoE) represents a significant advancement in the field of artificial intelligence, offering a powerful approach to scaling model capacity and improving computational efficiency. By dividing models into specialized expert networks and selectively activating them based on the input, MoE enables the development of larger and more capable models without a proportional increase in computational cost. While challenges remain, ongoing research and development efforts are continually expanding the capabilities and applications of MoE. MoE is considered by some to be the future of AI [2,7,10]. It is also being demystified for wider understanding [6,31]. Various resources explain MoE [15,29,43,44,50], and some provide simple explanations [**?**].

From fundamental architectures to cutting-edge applications, MoE has proven to be a versatile and powerful paradigm in machine learning. While challenges remain in areas like routing efficiency and training stability, ongoing research continues to push the boundaries of what's possible with expert-based models. As we look toward the future, MoE architectures are poised to play a central role in the development of more capable, efficient, and scalable AI systems [39,40,50].

Declaration

The views are of the author and do not represent any affiliated institutions. Work is done as a part of independent researcher. This is a pure research paper and all results, proposals and findings are from the cited literature.

References

1. Mixture of Experts (MoE) in AI Models Explained. https://blog.gopenai.com/mixture-of-experts-moe-in-ai-models-explained-2163335eaf85.
2. Mixture of Experts in AI: Boosting Efficiency. https://telnyx.com/learn-ai/mixture-of-experts.
3. Mixture of Experts (MoE) Explained. https://www.ultralytics.com/glossary/mixture-of-experts-moe.
4. Mixture of Experts (MoE): Unleashing the Power of AI. https://datasciencedojo.com/blog/mixture-of-experts/, 2024.
5. Mixture-of-Experts with Expert Choice Routing. https://research.google/blog/mixture-of-experts-with-expert-choice-routing/.
6. Demystifying Mixture of Experts (MoE): The Future for Deep GenAI Systems. https://blog.pangeanic.com/demystifying-mixture-of-experts-moe-the-future-for-deep-genai-systems.
7. Redefining AI with Mixture-of-Experts (MOE) Model. https://www.e2enetworks.com/blog/redefining-ai-with-mixture-of-experts-moe-model-mixtral-8x7b-and-switch-transformers.
8. Mixture of Expert Architecture. Definitions and Applications Included Google's Gemini and Mixtral 8x7B. https://ai.plainenglish.io/mixture-of-expert-architecture-7be02b74f311.
9. Neves, M.C. LLM Mixture of Experts Explained. https://www.tensorops.ai/post/what-is-mixture-of-experts-llm, 2024.
10. Mixture of Experts (MoE) Models: The Future of AI. https://www.linkedin.com/pulse/mixture-experts-moe-models-future-ai-saptashya-saha-buexc/.
11. Mixture of Experts(MoE) Revolutionizing AI with Specialized Intelligence. https://www.linkedin.com/pulse/mixture-expertsmoe-revolutionizing-ai-specialized-sanjeev-bora-jiuoc/.
12. Team, A.E. An Intro to Mixture of Experts and Ensembles, 2021.
13. Applying Mixture of Experts in LLM Architectures. https://developer.nvidia.com/blog/applying-mixture-of-experts-in-llm-architectures/, 2024.
14. Is GPT-4 a Mixture of Experts Model? Exploring MoE Architectures for Language Models. https://www.nownextlater.ai/Insights/post/is-gpt-4-a-mixture-of-experts-model-exploring-moe-architectures-for-language-models.

 doi:10.20944/preprints202505.1603.v1

15. Barr, A. Mixture-of-Experts Explained: Why 8 Smaller Models Are Better than 1 Gigantic One. https://alexandrabarr.beehiiv.com/p/mixture-of-experts, 2022.
16. All About Decentralized Mixture Of Experts (MoE): What It Is And Principles Of Operation. https://bullperks.com/all-about-decentralized-mixture-of-experts-moe-what-it-is-and-principles-of-operation/, 2024.
17. Jin, P.; Zhu, B.; Yuan, L.; Yan, S. MoE++: Accelerating Mixture-of-Experts Methods with Zero-Computation Experts. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2024.
18. Accelerate Mixtral 8x7B Pre-Training with Expert Parallelism on Amazon SageMaker. https://aws.amazon.com/blogs/machine-learning/accelerate-mixtral-8x7b-pre-training-with-expert-parallelism-on-amazon-sagemaker/.
19. DeepSeek Paper Offers New Details on How It Used 2,048 Nvidia Chips to Take on OpenAI. https://www.scmp.com/tech/big-tech/article/3310639/deepseek-paper-offers-new-details-how-it-used-2048-nvidia-chips-take-openai.
20. JIN. Mixture-of-Experts (MoE) Challenges: Overcoming Scaling and Efficiency Pitfalls, 2025.
21. How Do Mixture-of-Experts Layers Affect Transformer Models? https://stackoverflow.blog/2024/04/04/how-do-mixture-of-experts-layers-affect-transformer-models/, 2024.
22. Cerebras Launches World's Fastest Inference for Meta Llama 4. https://aijourn.com/cerebras-launches-worlds-fastest-inference-for-meta-llama-4/, 2025.
23. DeepSeek V3 0324 API, Providers, Stats. https://openrouter.ai/deepseek/deepseek-chat-v3-0324.
24. Shi, X.; Wang, S.; Nie, Y.; Li, D.; Ye, Z.; Wen, Q.; Jin, M. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2024.
25. Togootogtokh, E.; Klasen, C. TabularGRPO: Modern Mixture-Of-Experts Transformer with Group Relative Policy Optimization GRPO for Tabular Data Learning. *Qeios* **2025**. https://doi.org/10.32388/A9Q3VC.
26. Vision Language Models (Better, Faster, Stronger). https://huggingface.co/blog/vlms-2025, 2025.
27. Gupta, A. Forget ChatGPT? China's DeepSeek Is Working on Smarter, Self-Improving AI Models. https://www.livemint.com/technology/tech-news/forget-chatgpt-chinas-deepseek-is-working-on-smarter-self-improving-ai-models-11744017341248.html, 2025.
28. Vats, A. The Evolution of Mixture Of Experts: From Basics To Breakthroughs. https://pub.towardsai.net/the-evolution-of-mixture-of-experts-from-basics-to-breakthroughs-ab3e85fd64b3, 2024.
29. Mixture of Experts Explained. https://huggingface.co/blog/moe, 2025.
30. Understanding Mixture-of-Experts (MOE) in Large Language Models (LLMs) in Simple Terms. https://www.ctol.digital/news/mixture-of-experts-revolutionizing-llms/.
31. Zem, G. Explaining the Mixture-of-Experts (MoE) Architecture in Simple Terms, 2024.
32. Nayak, P. Create Your Own Mixture of Experts Model with Mergekit and Runpod. https://medium.aiplanet.com/create-your-own-mixture-of-experts-model-with-mergekit-and-runpod-8b3e91fb027a, 2024.
33. walidamamou. Mixture of Experts LLM & Mixture of Tokens Approaches-2024, 2024.
34. Sahoo, Juncheng Liu, T.A.C.L.C.X.D.X.L. Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts, 2024.
35. Mixture of Experts (MoE) for Financial. https://www.google.com/search?q=Mixture+of+Experts+(MoE)+for+financial.
36. Revolutionising Finance: How AI Is Transforming Investment and Risk Management, 2024.
37. Thompson (PhD), R. Can We Predict Market Moves Using MoE? https://medium.datadriveninvestor.com/can-we-predict-market-moves-using-moe-cafade516721, 2025.
38. Akira AI Unified Agentic AI Platform. https://www.akira.ai/.
39. Meta Hits Pause on Llama 4 Behemoth AI Model amid Capability Concerns.
40. Meta's Flagship AI Model Behemoth Delayed Release Raises Market Concerns. https://longportapp.com/en/news/240472785.
41. Alibaba Group Announces March Quarter 2025 and Fiscal Year 2025 Results. https://www.businesswire.com/news/home/20250514856295/en/Alibaba-Group-Announces-March-Quarter-2025-and-Fiscal-Year-2025-Results.
42. Nie, X. Codecaution/Awesome-Mixture-of-Experts-Papers, 2025.
43. Mixture of Experts. https://deepgram.com/ai-glossary/mixture-of-experts.
44. What Is Mixture of Experts? https://www.ibm.com/think/topics/mixture-of-experts, 2024.
45. Qwen2.5 VL! Qwen2.5 VL! Qwen2.5 VL! https://qwenlm.github.io/blog/qwen2.5-vl/.

 doi:10.20944/preprints202505.1603.v1

46. Ladd, V. Improving AI Data Privacy and Security Using MoE (Mixtures of Experts), 2023.
47. Torres, D.W. Mixture of Experts Models: Explained Simply, 2025.
48. walidamamou. Proficient Fine-Tuning via Mixture of Experts with PEFT, 2024.
49. CHOSUNBIZ. South Korea Initiates Feasibility Study for Advanced AGI Technology Development. https://biz.chosun.com/en/en-it/2025/03/05/6SWKUAXRCZAZ3DVRKZIL36Y4RQ/, 2025.
50. What Is a Mixture of Experts Model? https://www.itpro.com/technology/artificial-intelligence/what-is-a-mixture-of-experts-model, 2025.