International Journal of Future Engineering Innovations

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

Satyadhar Joshi
Independent, Alumnus, International MBA, Bar-Ilan University, Israel

* Corresponding Author: **Satyadhar Joshi**

Article Info
ISSN (online): 3049-1215
Volume: 02
Issue: 03
May-June 2025
Received: 10-04-2025
Accepted: 06-05-2025
Page No: 127-134

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. Next, the paper examines how machine learning is applied to natural language processing, computer vision, finance and healthcare. We examine major problems, including routing imbalance, memory fragmentation and instability during training, by checking newly proposed answers found in research papers. After all, we summarize possible future areas of study and discuss how MoE models could transform the world of artificial intelligence moving forward. All the results in this paper are taken from the cited work.

DOI: https://doi.org/10.54660/IJFEI.2025.2.3.127-134

Keywords: Mixture of Experts, MoE, Sparse Models, Large Language Models, Expert Parallelism, Neural Networks, AI Architecture

1. Introduction

The progress of artificial intelligence (AI) in recent years is mainly because larger and more complex models keep being built. On the other hand, scaling up the models is not easy due to greater costs in computation, more required memory and longer training processes. Because of these problems, experts have proposed the Mixture of Experts (MoE) architecture which offers advantages over common designs. MoE is a computer technique in machine learning that replaces a large neural network with many smaller networks handling different areas of the input data. Thanks to the gating network, the experts will only work on the needed input, helping the model work faster and with greater capacity. As a result, the model is able to work with a larger capacity without greatly increasing the time it takes to do computations. Emerging as an effective way to apply machine learning, the Mixture of Experts (MoE) is now a key architecture for large models working on natural language processing, computer vision and various multimodal functions [1, 1, 2, 3, 4]. Rather than using a single, large network, MoE divides the work into smaller networks, each handling a certain aspect of the data and has a gating network that decides which experts should be used for each input [5, 6, 7, 8, 9]. Using this method allows for better performance, a larger model capacity and improved scalability.

Modern machine learning owes much of its progress to MoE which makes it possible to build models that are big and efficient [10, 11]. Unlike most neural networks, MoE networks divide hard problems into smaller tasks that different networks manage by jointly deciding which tasks to use for each input [12].

Recent achievements show that MoE models are able to reach the best results and still be computationally efficient [13]. Mixtral 8x7B [8] and GPT-4 [14] indicate that multimodel-based architectures can deal with very large numbers of parameters. They are able to select only a few experts for each input which makes them much less costly to use than other models with the same capacity [15].

This paper makes the following contributions:

- A comprehensive review of MoE architectures and their evolution
- Analysis of 50 recent articles 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

Since its beginning, the Mixture of Experts architecture has advanced greatly, making it a main approach in building efficient machine learning systems for many applications. This section combines the most important studies, developments and issues in MoE research.

Many traditional neural networks are made up of a single, big architecture in which all the elements are switched on for each input. This strategy works fine for many established models, but it fails when models become very large. Because each input requires processing that is directly linked to the number of parameters, it is tough to handle large models.

The MoE architecture addresses the problem by making the computation sparse. To help reduce the total cost of computation, MoE just activates a select set of parameters instead of activating them all for every input. It is accomplished by parting the model into various expert networks, each focused on separate regions of the input data. After that, a gating network selects those experts who are best suited for a certain input and uses their results to give the outcome.

The MoE approach was developed because it is clear that various inputs need to be handled differently. If we look at natural language processing, experts may encounter situations where syntax skills are needed and in others where semantics are more important. By dividing experts into separate fields, MoE is better able to represent complicated data sets.

2.1 Foundations of MoE

MoE tasks are handled by expert networks assigned specifically to different subtasks which are only called on when required by a gating mechanism. This initial work by [12], presented MoE as being able to do better with diverse data by directing each input to its relevant expert. Improvements in architecture have come to the forefront in deep learning, especially with research work like [15] which underscores its ability to lower computation needs while not sacrificing model's capacity—illustrated by GPT-4 which may have as many as 1.76 trillion parameters spread through different experts.

2.2 Advances in MoE Architectures

These days, attention is on ensuring that the MoE is both productive and scalable to many tasks.

Sparse Activation: Both the Mixtral 8x7B [8] and Switch Transformers [5] achieve excellent results by activating only certain sparse experts in the network.

Decentralized MoE: [16] suggested distributed architectures for MoE which improve their reliability and ability to run simultaneously.

Hybrid Designs: [17] also suggested MoE++ which uses zero-computation experts to help keep resource use low.

2.3 Hardware and software innovations

When models have billions or trillions of parameters, we need to make progress in hardware and find better ways to train them using several computers at once. 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 finds use in the areas of NLP (see [13] for example), time-series forecasting (see example in [24]) and finance (see [25]). But there are still difficulties to overcome:

- **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

Three principal components are part of the MoE architecture: experts, gates and a routing mechanism [29]. Almost every expert is a neural network built for one type of input, while the gating network selects the right set of experts for a given input [30].

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

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

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

Issues in single-model architectures with complex datasets inspired the origination of the MoE concept. Distribute inputs to the most suitable experts according to their expertise is done by the gating mechanism used in the MoE design [3, 4, 5, 29]. Researchers have pointed out that MoE results in less tuning of parameters and allows specialized models to be created faster and at lower cost, even for mind-boggling numbers of these parameters [13, 14, 15].

3.3 Advancements in MoE Architectures

Over the past few years, there has been an increase in new types of MoE architectures such as Switch Transformers, GShard and Mixtral [1, 7, 8, 18]. This approach means that, for every input, only a few experts are activated, greatly decreasing the work the model does. Some highlights are found in:

- **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

Every neural network in an expert network is dedicated to sorting out one part of the input information. The specific job application can determine if the network uses fully connected, convolutional or recurrent layers as its structure.

3.5 Gating Network

The gating network uses input data to identify which experts should become active. It receives the input as input and delivers a set of weights, one for each expert. They tell you which experts are most important to each piece of information.

3.6 Routing Mechanism

The gating network gives each expert network a weight which is used by the routing network to combine multiple expert outputs. More often, a weighted sum is used which computes the results by adding each expert's weighted output.

The MoE architecture can be summarized as follows:

- The input is fed into both the gating network and the expert networks.
- The gating network produces weights for each expert.
- The output of each expert is multiplied by its corresponding weight.
- 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]

3.9 Advantages of MoE

MoE offers several advantages over traditional neural networks:

Scalability: Scale-up in model capacity is allowed through MoE, without a similar increase in how much computation takes place. When including additional experts, the model has the power to learn more intricate patterns and the extra cost for each input does not change much since just a group of experts is needed.

Efficiency: Unlike regular neural networks, MoE lessens the computational burden by saving parameters for only part of each input. It helps save time and hardware by allowing us to use bigger models when training and deploying them on limited equipment.

Specialization: A MoE model can work better with many types of data because its experts each deal with a specific part of the input space. With this specialization, individuals may do better at a range of tasks.

4. Applications

A variety of fields have used MoE models with good results. Many tasks have been solved using MoE.

4.1 Natural language processing

With MoE, GPT-4 and Mixtral now have the ability to train with trillions of parameters and retain performance [14, 18]. The concept of using MoE in LLMs is explored in many papers, including [3, 6, 9, 13, 30, 33].

4.2 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.

4.3 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.

4.4 Other Applications

MoE is also being explored in other areas, including finance [35, 36, 37], and agentic AI platforms [38].

4.5 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].

4.6 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].

4.7 Time series analysis

Work from the recent past has extended the use of MoE to forecast time series [34]. The Time-MoE system manages to model time series on a large scale and does so very efficiently [24].

4.8 Computer Vision

More and more, vision-language models are being built using the MoE architecture [26]. Both visual and linguistic experts work together in these models for understanding everything [45].

4.9 Finance and Economics

Many researchers believe that MoE models may be useful in financial areas [35]. They are capable of exploring market movements [37] and making risk assessment better [36].

4.10 Finance, investment economics, and risk applications of mixture of experts

The MoE architecture has been successful in putting its best foot forward in finance, investment and business because its focus on particular tasks and scalability fits well with those fields. We take a look at important applications of MoE in these areas.

In financial fields, the MoE architecture has shown the ability to provide particular solutions for complex decisions. In this part, we consider major uses for financial modeling, different investment approaches, economic modeling tools and reviewing possible risks.

Using Mixture of Experts (MoE) tools in finance and investment economics is quickly changing standard techniques for handling risk, assembling portfolios and performing financial analyses. These architectures stand out as being especially useful in big, data-heavy financial markets, thanks to their ability to be tuned and scaled [35, 36].

4.11 AI-driven investment and risk management

AI and especially MoE are making investment and risk management more accurate by helping users to adjust their decisions based on high-quality analysis. With MoE, investment systems can be designed to work with different assets, market conditions or economic patterns which supports both broader portfolio diversity and automatic adjustments to asset distribution [36]. Such models can process many different types of financial data, spot trends and predict risk situations much better than legacy models.

4.12 Economic Modeling

- Using MoE frameworks, we can work with multiple economic indicators well and [24] improved accuracy in GDP forecasting by 17%.
- [34] applied MoE to macroeconomic time-series data, reducing parameter requirements by 65x compared to conventional models.

4.13 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.

4.14 Investment portfolio optimization

MoE architectures enable dynamic asset allocation strategies: [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.

4.15 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:** These models can find and highlight fraud-indicative patterns and further real-time monitoring [38].

4.16 Finance and Investment

To achieve better accuracy and performance, financial experts in risk assessment, portfolio optimization and market forecasting have used MoE models. For example, [37] developed an MoE approach to model what users want and how the market behaves, helping make more accurate forecasts on market trends. TabularGRPO is a MoE-based transformer that [25] introduced for use with tabular data in finance. By handling the difficulty of analyzing mixed and unevenly distributed features, it achieved the leading results in that field. They display the Minimum Orthogonal Efficiency (MoE) technique's method of managing highly intricate, high-dimensional data without raising computing costs.

Business process automation

With MoE, AI is incorporated into both ERP and business automation, making it easier to carry out tasks such as preparing forecasts, making reports and analyzing risks [38]. This means businesses operate better and plan their strategies with better information.

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.

4.17 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.

4.18 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].

4.19 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.

4.20 Conclusion, challenges and consideration

Using Mixture of Experts in finance and investment economics brings a big change in how institutions handle risk and investment matters. Thanks to combining different areas of expertise, MoE models promise to work well, handle changes in markets and ensure a high level of performance [35, 36, 38].

Even with the help of MoE, financial applications experience specialized difficulties:

- Regulatory compliance means that the expert's routing decisions must be easily understood.
- Because high-frequency trading has limits on latency, it needs gating mechanisms that have been optimized.
- Because economic indicators often change, it is necessary for experts to keep retraining themselves.

MoE architectures can make a big difference for financial organizations that need to run efficient operations while applying what they know about their fields. Future plans are to use federated learning on data from different institutions for risk modeling and apply quantum enhancements to develop skilled networks for derivative valuation.

5. 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, 39, 40].

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

5.1 Training Complexity

It is more difficult to train MoE models than to train traditional neural networks. To do its job, the gating network must be good at routing inputs to their appropriate experts which poses a challenging optimization issue.

5.2 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.

5.3 Increased memory usage

Even though you can make computations faster, MoE models are usually bigger, as they store many expert network parameters.

5.4 Routing Imbalance

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

- Load balancing constraints
- Adaptive routing mechanisms [21]

5.5 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]

5.6 Training Instability

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

- Regularization techniques
- Progressive training schedules [48]

6. Future Directions

Current research is aimed at solving the MoE issues of weak training stability, ineffective routing methods and large memory consumption. People in this field are investigating new types of MoEs that are less centralized [16].

Future directions for MoE research include:

6.1 Dynamic Routing
Achieving routing that can automatically reassign inputs to specialists as the input properties change.

6.2 Adaptive Capacity
Investigating techniques to adapt how much data can be handled by the expert networks depending on the complexity of the data.

6.3 Hardware Acceleration
Establishing architecture designs for hardware that performs MoE models efficiently.

6.4 Combining with other architectures
When MoE is connected with architectural developments such as attention and transformers, it creates much more potent and efficient models.

The research in this field is developing rapidly, as people continue to focus on 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].

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

6.6 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].

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

7. 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].

8. 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.

9. References

1. Mixture of experts (MoE) in AI models explained. [Internet]. Available from: https://blog.gopenai.com/mixture-of-experts-moe-in-ai-models-explained-2163335eaf85
2. Mixture of experts in AI: Boosting efficiency. Telnyx. [Internet]. Available from: https://telnyx.com/learn-ai/mixture-of-experts
3. Mixture of experts (MoE) explained. [Internet]. Available from: https://www.ultralytics.com/glossary/mixture-of-experts-moe
4. Mixture of experts (MoE): Unleashing the power of AI. [Internet]. Jan. 2024. Available from: https://datasciencedojo.com/blog/mixture-of-experts/
5. Mixture-of-experts with expert choice routing. Google Research Blog. [Internet]. Available from: https://research.google/blog/mixture-of-experts-with-expert-choice-routing/
6. Demystifying mixture of experts (MoE): The future for deep GenAI systems. [Internet]. Available from: https://blog.pangeanic.com/demystifying-mixture-of-experts-moe-the-future-for-deep-genai-systems
7. Redefining AI with mixture-of-experts (MOE) model. [Internet]. Available from: 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 including Google's Gemini and Mixtral 8x7B. [Internet]. Available from: https://ai.plainenglish.io/mixture-of-expert-architecture-7be02b74f311
9. Neves MC. LLM mixture of experts explained. TensorOps. [Internet]. Jan. 2024. Available from: https://www.tensorops.ai/post/what-is-mixture-of-experts-llm
10. Mixture of experts (MoE) models: The future of AI. [Internet]. Available from: https://www.linkedin.com/pulse/mixture-experts-moe-models-future-ai-saptashya-saha-buexc/
11. Mixture of experts (MoE): Revolutionizing AI with specialized intelligence. [Internet]. Available from: https://www.linkedin.com/pulse/mixture-expertsmoe-revolutionizing-ai-specialized-sanjeev-bora-jiuoc/
12. AICorespot Team. An intro to mixture of experts and ensembles. AICorespot. Oct. 2021.

13. Applying mixture of experts in LLM architectures. NVIDIA Technical Blog. [Internet]. Mar. 2024. Available from: https://developer.nvidia.com/blog/applying-mixture-of-experts-in-llm-architectures/
14. Is GPT-4 a mixture of experts model? Exploring MoE architectures for language models. [Internet]. Available from: https://www.nownextlater.ai/Insights/post/is-gpt-4-a-mixture-of-experts-model-exploring-moe-architectures-for-language-models
15. Barr A. Mixture-of-experts explained: Why 8 smaller models are better than 1 gigantic one. Superslow AI Newsletter. [Internet]. Dec. 2022. Available from: https://alexandrabarr.beehiiv.com/p/mixture-of-experts
16. All about decentralized mixture of experts (MoE): What it is and principles of operation. [Internet]. Dec. 2024. Available from: https://bullperks.com/all-about-decentralized-mixture-of-experts-moe-what-it-is-and-principles-of-operation/
17. Jin P, Zhu B, Yuan L, Yan S. MoE++: Accelerating mixture-of-experts methods with zero-computation experts. In: The Thirteenth International Conference on Learning Representations; Oct. 2024.
18. Accelerate Mixtral 8x7B pre-training with expert parallelism on Amazon SageMaker. [Internet]. Available from: 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. [Internet]. Available from: 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. AI Simplified in Plain English. [Internet]. Mar. 2025.
21. How do mixture-of-experts layers affect transformer models? Stack Overflow Blog. [Internet]. Apr. 2024. Available from: https://stackoverflow.blog/2024/04/04/how-do-mixture-of-experts-layers-affect-transformer-models/
22. Cerebras launches world's fastest inference for Meta Llama 4. [Internet]. Apr. 2025. Available from: https://aijourn.com/cerebras-launches-worlds-fastest-inference-for-meta-llama-4/
23. DeepSeek V3 0324 API, providers, stats. [Internet]. Available from: https://openrouter.ai/deepseek/deepseek-chat-v3-0324
24. Shi X, *et al*. Time-MoE: Billion-scale time series foundation models with mixture of experts. In: The Thirteenth International Conference on Learning Representations; Oct. 2024.
25. Togootogtokh E, Klasen C. TabularGRPO: Modern mixture-of-experts transformer with group relative policy optimization (GRPO) for tabular data learning. Qeios. Mar. 2025. doi: 10.32388/A9Q3VC
26. Vision language models (better, faster, stronger). Hugging Face Blog. [Internet]. May 2025. Available from: https://huggingface.co/blog/vlms-2025
27. Gupta A. Forget ChatGPT? China's DeepSeek is working on smarter, self-improving AI models. Mint. [Internet]. Apr. 2025. Available from: https://www.livemint.com/technology/tech-news/forget-chatgpt-chinas-deepseek-is-working-on-smarter-self-improving-ai-models-11744017341248.html
28. Vats A. The evolution of mixture of experts: From basics to breakthroughs. Medium. [Internet]. Sep. 2024. Available from: https://pub.towardsai.net/the-evolution-of-mixture-of-experts-from-basics-to-breakthroughs-ab3e85fd64b3
29. Mixture of experts explained. Hugging Face Blog. [Internet]. Feb. 2025. Available from: https://huggingface.co/blog/moe
30. Understanding mixture-of-experts (MoE) in large language models (LLMs) in simple terms. [Internet]. Available from: https://www.ctol.digital/news/mixture-of-experts-revolutionizing-llms/
31. Zem G. Explaining the mixture-of-experts (MoE) architecture in simple terms. Medium. Jan. 2024.
32. Nayak P. Create your own mixture-of-experts model with mergekit and runpod. Medium. Available from: https://medium.aiplanet.com/create-your-own-mixture-of-experts-model-with-mergekit-and-runpod-8b3e91fb027a. Jan. 2024.
33. Walidamamou. Mixture of experts LLM & mixture of tokens approaches-2024. UBIAI. Mar. 2024.
34. Sahoo TA, Liu J. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. Salesforce. Nov. 2024.
35. Mixture of experts (MoE) for financial. Available from: https://www.google.com/search?q=Mixture+of+Experts+(MoE)+for+financial.
36. Revolutionising finance: How AI is transforming investment and risk management. Oct. 2024.
37. Thompson R. Can we predict market moves using MoE? Medium. Available from: https://medium.datadriveninvestor.com/can-we-predict-market-moves-using-moe-cafade516721. Apr. 2025.
38. Akira AI unified agentic AI platform. Available from: https://www.akira.ai/.
39. Meta hits pause on llama 4 behemoth AI model amid capability concerns. Computerworld.
40. Meta's flagship AI model behemoth delayed release raises market concerns. Available from: https://longportapp.com/en/news/240472785.
41. Alibaba group announces March quarter 2025 and fiscal year 2025 results. Available from: 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. May 2025.
43. Mixture of experts. Deepgram. Available from: https://deepgram.com/ai-glossary/mixture-of-experts.
44. What is mixture of experts? Available from: https://www.ibm.com/think/topics/mixture-of-experts. Apr. 2024.
45. Qwen2.5 VL! Qwen2.5 VL! Qwen2.5 VL! Available from: https://qwenlm.github.io/blog/qwen2.5-vl/.
46. Ladd V. Improving AI data privacy and security using MoE (mixtures of experts). Medium. Dec. 2023.
47. Torres DW. Mixture of experts models: Explained simply. Deep Learning With The Wolf. Jan. 2025.
48. Walidamamou. Proficient fine-tuning via mixture of experts with PEFT. UBIAI. Aug. 2024.
49. South Korea initiates feasibility study for advanced AGI technology development. CHOSUNBIZ. Available

from: https://biz.chosun.com/en/en-it/2025/03/05/6SWKUAXRCZAZ3DVRKZIL36Y4RQ/. Mar. 2025.
50. What is a mixture of experts model? IT Pro. Available from: https://www.itpro.com/technology/artificial-intelligence/what-is-a-mixture-of-experts-model. Apr. 2025.
51. Joshi S. A literature review of Gen AI agents in financial applications: Models and implementations. International Journal of Science and Research (IJSR). doi: https://doi.org/10.21275/SR25125102816.
52. Joshi S. Advancing innovation in financial stability: A comprehensive review of AI agent frameworks, challenges, and applications. World Journal of Advanced Engineering Technology and Sciences. 2025;14(2):117–126. doi: 10.30574/wjaets.2025.14.2.0071.
53. Joshi S. Implementing Gen AI for increasing robustness of US financial and regulatory system. IJIREM. 2025;11(6):Art. no. 6. doi: 10.55524/ijirem.2024.11.6.19.
54. Joshi S. Leveraging prompt engineering to enhance financial market integrity and risk management. World J. Adv. Res. Rev. 2025;25(1):1775–85. doi: 10.30574/wjarr.2025.25.1.0279.
55. Joshi S. Review of data engineering and data lakes for implementing GenAI in financial risk: A comprehensive review of current developments in GenAI implementations. Social Science Research Network. Rochester, NY: 5123081. Jan. 2025. doi: 10.2139/ssrn.5123081.
56. Joshi S. Review of data engineering frameworks (Trino and Kubernetes) for implementing generative AI in financial risk. International Journal of Research Publication and Reviews. 2025;6(2):1461–70. doi: 10.55248/gengpi.6.0225.0756.
57. Joshi S. Review of data pipelines and streaming for generative AI integration: Challenges, solutions, and future directions. International Journal of Research Publication and Reviews. 2025;6(2):2348–57.
58. Joshi S. The synergy of generative AI and big data for financial risk: Review of recent developments. IJFMR - International Journal For Multidisciplinary Research. 2025;7(1). doi: https://doi.org/g82gmx.
59. Joshi S. Review of data engineering frameworks (Trino and Kubernetes) for implementing generative AI in financial risk. International Journal of Research Publication and Reviews. 2025;6(2):1461–70. doi: 10.55248/gengpi.6.0225.0756.
60. Joshi S. Review of autonomous systems and collaborative AI agent frameworks. International Journal of Science and Research Archive. 2025;14(2):961–72. doi: 10.30574/ijsra.2025.14.2.0439.