WISDOM ACADEMIC

# Data Ethical and CyberSecurity

ISSN: 3104-6479 | E-ISSN: 3104-6487

*Volume 1 , Issue 3*

**Article**

# The Research on Algorithms, Algorithm Optimization, and Applications of Large-Scale AI Models

Xiang Zhang [1,2], Hazirah Bee Yusof Ali[1,*], Qi Xi[2]

[1]Faculty of Information Technology, City University Malaysia, Kuala Lumpur 46100, Malaysia.

[2]School of Information Engineering, Jingdezhen University, Jingdezhen 333000, China.

[*]**Corresponding author:** Hazirah Bee Yusof Ali, hazirah.bee@city.edu.my.

**CITATION**

Zhang X, Yusof Ali HB & Xi Q. The Research on Algorithms, Algorithm Optimization, and Applications of Large-Scale AI Models. Data Ethical and CyberSecurity. 2025; 1(3): 262.

https://doi.org/10.63808/decs.v1i3.262

**ARTICLE INFO**

Received: 14 November 2025
Accepted: 21 November 2025
Available online: 22 November 2025



**Abstract:** In recent years, artificial intelligence systems, such as Large Language Models (LLMs) and multimodal large models, have developed rapidly. With the exponential increase in model parameters, computational complexity, algorithm designs, and engineering implementations face incomparable challenges. How to train models with parameters in the range of hundred billion to trillion with restricted computing capacity and storage resources, efficient fine-tuning, deployment, further latency reduction, and decreased memory footprint have become the most fundamental problems in modern studies. Benefiting from the systematic analysis on respective reviews, different papers, and studies in these years, based on above analysis, in this article, there would be analysis on three different aspects, separately: Firstly, at algorithm foundations, there's investigation on evolution in optimizers, network architectures, and training patterns. Secondly, in algorithm levels, there's summary on core areas like memory, resource, efficient parameter fine-tuning, model compression, and other directions. Thirdly, in levels concerning applications, problems, challenges, there's observation on applications, their respective feasibilities, limitations in modern models with optimized algorithms. By comparison on different studies on these areas, there's illustration on trends in algorithm optimizations for large-scale AI models, identification on key problems in modern studies, and illustration on

different possibilities in new studies. This article looks for providing a systematic literature analysis for researchers, engineers, participants in studies concerning algorithms in large models, optimizations, in AI studies, for necessary references in their studies.

**Keywords:** algorithm optimization; large-scale model; distributed training; efficient fine-tuning of parameters; memory optimization

## 1. Introduction

Computing power is improving at such a speed, with increasing accumulation of large-scale data resources, which leads to rapid development in artificial intelligence technology. Among these, models with the Transformer-based structure have played an increasingly crucial role in being the core support for modern-scale artificial intelligence systems. The innovation in model structure is to break through limitations in context relationships on shorter sequences, with extreme scalability, thus accelerating the appearance of a series of models with scales from hundreds of millions to trillions, such as the GPT series, PaLM, and M6 models (Tu et al., 2024). These models have showcased superb performance with generalization in natural language understanding, image synthesis, multimodal inference, and other areas, becoming an important driving force for artificial intelligence to move towards general intelligence.

On the other hand, with the continually expanding size of the model, the problems caused by it during the computational process have also attracted increasing attention. On one hand, the training process for large models is extremely computationally intensive, requiring numerous high-performance GPUs/TPUs to work in parallel, with training cycles taking weeks or months. On the other hand, with such explosive growth in model parameters, training is not only computationally intensive, with high memory utilization, communication costs, and resulting limitations on model scalability, along with other constraints on model training, such limitations on model scalability have also restricted model training. Again, for inference/reasoning, with such large model sizes, problems such as inference latency, energy, and costs have restricted its large-scale applications in real-world applications (Bai et al., 2024; Tian et al., 2025). Hence, optimizing computational costs while

maintaining model accuracy has become an important common concern for experts in these fields.

With such context, the studies on “algorithms” and “algorithm optimization” have begun to transform from theory to practice, becoming a double fulcrum to facilitate large model development. Algorithmic innovations are primarily centered on optimizers, regularization techniques, innovation in network architectures, and paradigm evolution in training algorithms. For example, the paradigm shifts from traditional SGD (Stochastic Gradient Descent) to modern optimizers like AdamW, LAMB, etc., now allows for efficient convergence in large parameter spaces. Along with these advancements, there are also improved variants of the Transformer model being developed periodically, such as Sparse Transformers, Performer, etc., hoping to increase computational efficiency while maintaining model capacity.

Conversely, the emphasis on algorithm optimization studies is on problems at the resource level, especially on training model algorithms in low computing power and memory environments. Commonly, these areas consist of minimizing video memory, gradient checkpoint technology, parallel algorithms for training in distributed environments (data parallelism, model parallelism, pipeline parallelism, etc.), efficient model parameter tuning (LoRA, prefide-tuning, etc.), model compression, model distillation, among others. These have significantly eased the computational burdens posed by large-scale models on different levels, thereby allowing models initially requiring supercomputer levels for support to execute in a general computing power environment.

On the basis of core reviews and research reports in recent years, based on the method of literature analysis, this paper undertakes a systematic analysis, trying to carve out a complete logical chain from algorithm theory foundation to in-depth analysis on optimization strategies in the engineering level, with an end goal reaching towards practice and trends in applications. More specifically, this paper will firstly examine algorithm foundations for large-scale models in AI, such as optimizers, regularization, and network architectures, secondly, it would comprehensively analyze optimization in resource-scarce scenarios in recent years, such as video memory, computing, training architectures, efficient fine-tuning, model compression, distillation, and lastly, based on core applications for typical large-scale models, performance analysis for these technologies on application systems in practice, to

finally explore research directions in algorithm collaboration, sustainable training, intelligent optimization, and the like in upcoming years.

On the whole, in this paper, efforts will be made to explore the development process and trends of large-scale AI model algorithm studies in the analysis of related literature, as well as clarify the crucial importance of algorithm optimization in realizing highly efficient intelligent systems. The references in this paper include most of the typical studies in the literature on related fields, such as (Bottou, 2018; He, 2021; Shen, 2023; Tu, 2024, etc.), which have high representativity with reference value, acting as evidence to support analysis in next chapters.

## 2. Method

This article employs the method of systematic literature analysis, endeavoring to guarantee the representativeness of literature sources, coupled with the objectivity of conclusions regarding research on large-scale AI models in recent years.

On the one hand, for the literature retrieval process, multiple databases and conference papers with authoritative academics, advanced technology, such as arXiv, Springer, MDPI, IEEE Xplore, ACM Digital Library, together with conference papers on artificial intelligence and machine learning, have been chosen for this paper. Among these keywords, those included in the literature searches primarily consist of “large-scale models,” “algorithm optimization,” “distributed training,” “parameter-efficient fine-tuning,” “memory-efficient training,” while other papers pertain to some core directions about algorithm designs/optimizations for large-scale models. It is necessary to emphasize, in conducting these studies to guarantee their timeliness, there should be particular concern regarding reviews, papers with high citations, or those viewed with prominence in the researchers’ community.

Secondly, in the process of screening and organizing literature, to delve into the research topic, I divide the literature into four categories: (1) Algorithm foundation, with regards to optimizers, network topology, training modes, etc., to understand their evolution process; (2) Memory 和 distributed optimization, with reference to optimize computing efficiency in scenarios with restricted resources; (3) Parameter fine-tuning, emphasizing efficient training techniques like LoRA, Prefide-tuning, Adapter; (4) Application Evaluation: Investigate the training, fine-tuning, and implementation performance, as well as technical difficulties in actual applications, of these

algorithms in model training, fine-tuning, and implementation. These four kinds of literature have been sub-categorized in an attempt to sum up their essence, proposed technology, conclusions, strengths, and weaknesses.

Finally, in the analysis stage, in this paper, there is a horizontal and vertical comparison with different research results, aiming to explore innovation in theory, engineering feasibility, and application adaptability. By cross-comparison with different literatures, in this paper, an attempt is made to explore different key issues, which have not yet received due consideration in previous studies, and accordingly, different directions for future studies shall be determined.

It needs to be emphasized, however, that there is no new experimental verification or analysis in this paper. Its result solely relies on its systematic analysis and logical inference from extant literature. The benefit of adopting such an approach primarily resides in its capacity to capture the general vision from multiple angles, acquire a macroscopic concept on algorithms regarding large-scale AI models, thereby creating a solid foundation for innovation in theory based on literature.

# 3. Results

## 3.1. Algorithm Foundation: Optimizer, Training Paradigm and Network Structure

Classic optimizers like SGD and Adam are still at the foundation of deep network training, yet for training with parameters on a massive large-scale, traditional training methods will face new challenges in terms of stability, convergence rate, and generalization performance (Bottou, 2018). Reviews like He et al. (2021) highlight not only breakthroughs in optimizers, such as those proposed by LAMB et al. for high-volume training, but also the importance of other aspects like scheduling methods for learning rates, regularization techniques like SAM, and gradient noise. On the network structure, while there is no doubt about the popularity in large-scale models like Transformer, its self-attention mechanism has provided what is generally considered to be optimal parallelization properties for dependence on long-range patterns. To alleviate complexity, the researchers adopted sparse attention or optimized long sequences (Tu et al., 2024).

Additionally, algorithmic concepts like meta-learning, AutoML, and low-rank matrix decomposition have also contributed to improving the aptness of large-scale models in adapting to new tasks (Hospedales et al., 2020). On the whole, “algorithmic foundations” research areas have shifted from loss functions to optimize “stability, generalization, and resource controllability” together.

## 3.2. Algorithm Optimization: Engineering Methods for Scale

Commonly, in literature, optimization techniques for large-scale models have been grouped into several categories, including memory/resource optimization, parallelization techniques, parameter fine-tuning techniques (PEFT), and compression/acceleration techniques at the inference level.

(1) Memory and Resource Optimization

Tian et al. (2025) and Bai et al. (2024) provide a comprehensive overview regarding the optimization of memory efficiency. These crucial technologies include mixed precision training, FP16/BF16, activation checkpointing, off-loading optimizer status, and parameters, along with hierarchical storage, among other things. Mixed precision training, checkpointing, off-loading, or other techniques enable the training of large models with available hardware resources, through computing for space trade-off in terms of time, or optimized data representations to reduce peak memory (Tian et al., 2025).

Mixing accuracy could not only save video memory, but it could also improve throughput. The checkpointing portion decreases forward storage by re-computing activations. Though off-loading shifts some information from small, fast storage to slow, larger storage, these four techniques are basically employed together to create an efficient memory management plan (Mei, et al., 2024; Liu, et al., 2024).

(2) Distributed training, parallel methods

Distributed parallelism: Distributed parallelism encompasses data parallelism, tensor parallelism, model parallelism, pipeline parallelism, and their hybrids. Communication efficiency, balance, and fault-tolerant techniques have been discussed in their works by Tang et al. (2020) and Zeng et al. (2025). It has been proved from their experiences that for hybrid parallelism, in which tensor-parallelism is combined with pipeline-parallelism, there are lower memory requirements for individual devices for efficient training, with stricter requirements for communication bandwidth, synchronization, and error correction (Tang et al., 2020; Zeng et al., 2025).

On the other hand, communication compression (quantization of the gradient), late updates (asynchrony/stale updates), or hierarchical synchronization mechanisms could alleviate communication bottlenecks, especially in networks with high latency or training across multiple data centers.

(3) Parameter Fine-tuning

To enable pre-trained large-scale models to be scalable, deployable, and efficient in different downstream tasks, PEFT technology has received intensive research efforts in the past few years. Some popular techniques belong to: Adapters, Prompt Tuning, LoRA (Low-Rank Adaptation), and sparse fine-tuning, etc. The underlying concept for these approaches is to fine-tune only a small quantity of new parameters or patches with low ranks, while maintaining other model parameters fixed, in order to greatly save resources in terms of video memory and storage, as well as simplify model management complexity. By real-world practice, it's observed that for most downstream tasks, there is nearly equal performance but significantly lower costs for PEFT compared to full fine-tuning.

(4) Reasoning optimization and model compression

During reasoning, latency, throughput, and energy are aimed to be optimized. These methods include: Quantization, Pruning, Data distillation, orSparse activation in Hybrid-Expert architectures (MoE). It has been observed by Bai et al. (2024); Liu et al. (2024) that model compression leads to substantial inference costs in many scenarios, while maintaining its accuracy, yet these compression algorithms have to balance these interactions in different scenarios such as edge devices, cloud, or something similar.

## 3.3. Applications and Challenges of Large-scale AI Models

Tu et al. (2024) have undertaken an extensive survey on the evolution, utilization, and applications of LLM, LVM, and LMM. Large models have proved to have immense applications in areas such as text generation, image analysis, multimodal retrievals, recommendation, and high-performance computing (Tu et al., 2024). But it has also been reiterated in several studies that there are some really tough problems:

Cost and Sustainability: The energy use in training, maintaining, and operating large-scale models has received increasing social concern, resulting in green computing becoming an inevitable research direction in society (Bai et al., 2024).

Interpretability and Security: It is difficult to interpret the decision paths in large models, and there could be security concerns about bias (Shen et al., 2023).

Vertical Domain Adaptation: Generalized large models have to be adapted for vertical applications with high demand, such as healthcare and finance. PEFT and efficient training data play an important role in this process.

Multimodal consistency: For LMM, there remain some problems in cross-modal semantic alignment and consistency generation (Tu et al., 2024).

## 4. Discussion

On the basis of the current literature, several conclusions may be deduced: Firstly, algorithmic progress must proceed together with algorithmic engineering. It is not enough to simply have an improved optimizer or a better neural network structure, which is not sufficient to cope with such problems in large-scale training. On the other hand, to significantly improve generalization and efficiency problems, parallelism in engineering is not enough (Bottou, 2018; Tang et al., 2020). Secondly, there exists an efficient fine-tuning and compression method for parameters, which offers a feasible solution to incorporating large models to be shared for adoption by those institutions with less powerful resources too. Thirdly, there needs to be increased cross-layer cooperation in algorithm designs, based on considerations for parallelism and friendliness for memory in algorithm designs, with consideration for convergence in algorithm designs for parallelism (Zeng et al., 2025).

Moreover, some other promising areas for further research work include: Automated parallel strategies (automatic definition of tensor/model/data-parallel splits), more efficient fine-tuning frameworks (robust adaptability with minimized samples needed), Green training frameworks (energy efficiency, carbon footprint, included among primary objectives for optimization), as well as training/fine-tuning frameworks with improved interpretability/privacy preservation (Bai, et al., 2024).

## 5. Conclusion

On the basis of some significant reviews and tech reports in these years, in this paper, we comprehensively discuss and analyze algorithm foundations for, as well as

optimization techniques related to, large-scale AI models. In general, with regards to large-scale models, there is now a co-evolution process from mixed precision to activation checkpointing, to distributed parallelism, to PEFT, aiming to strike "performance-resources-deploy ability." Future contributions in this direction demand closer cross-layer cooperation, sustainable training, and readiness for vertical applications. For those researchers willing to engage in such an area, there are two directions in cross-cutting: "algorithm for resource efficiency" and "system for deployment."

**Conflict of interest:** The authors declare no conflict of interest.

**Funding:** This research received no external funding.

## References

[1] Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Yang, C., Cheng, Y., & Zhao, L. (2023). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv*. https://doi.org/10.48550/arXiv.2401.00625

[2] Bottou, L. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223–311.

[3] He, X., Xue, F., Ren, X., & You, Y. (2021). Large-scale deep learning optimizations: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.2111.00856

[4] Liu, J., Tang, P., Wang, W., Ren, Y., Hou, X., Guo, M., & Li, C. (2024). A survey on inference optimization techniques for mixture of experts models. *arXiv*. https://doi.org/10.48550/arXiv.2412.14219

[5] Mei, T., Zi, Y., Cheng, X., Gao, Z., Wang, Q., & Yang, H. (2024). Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks. *arXiv*. https://doi.org/10.48550/arXiv.2405.11704

[6] Shen, L., Sun, Y., Yu, Z., Ding, L., Tian, X., & Tao, D. (2023). On efficient training of large-scale deep learning models: A literature review. *arXiv*. https://doi.org/10.48550/arXiv.2304.03589

[7] Tang, Z., Shi, S., Wang, W., Li, B., & Chu, X. (2020). Communication-efficient distributed deep learning: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.2003.06307

[8] Tian, X., Li, Y., Zhang, Q., & Zhou, W. (2025). A survey on memory-efficient large-scale model training. *arXiv*. https://doi.org/10.48550/arXiv.2501.11847

[9] Tu, X., He, Z., Huang, Y., Zhang, Z.-H., Yang, M., & Zhao, J. (2024). An overview of large AI models and their applications. *Visual Intelligence*, 2, Article 34.

[10] Zeng, F., Li, X., Wang, Y., Chen, H., & Zhao, Q. (2025). Distributed training of large language models: A survey. *Information Systems*. Advance online publication.