Challenges and Strategies in the Development of Large Models

1st Ziqi Wang
College of Computer Science
Beijing University of Technology
Beijing, China
ziqi_wang@emails.bjut.edu.cn

2nd Jing Bi
College of Computer Science
Beijing University of Technology
Beijing, China
bijing@bjut.edu.cn

3rd Hailiang Zhao
School of Software Technology
Zhejiang University
Ningbo, China
hliangzhao@zju.edu.cn

4th MengChu Zhou
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, USA
zhou@njit.edu

2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS) | 979-8-3315-2164-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICHMS65439.2025.11154190

***Abstract*—In recent years, the rapid advancement of Large Models (LMs), including large language models, visual foundation models, and multimodal LMs, has significantly transformed various fields. These models are evolving at a very fast pace, and their development has benefited numerous industries considerably. However, inherent architectural limitations in LMs, such as hallucinations and challenges in error localization, restrict their potential. Addressing these issues effectively is crucial for their continued progress. This work provides an overview of the evolution of LMs and highlights key challenges they face, including high data and energy demands, catastrophic forgetting, limited reasoning capabilities, and fault localization. Strategies to mitigate these challenges are proposed, followed by a discussion on applying LMs in smart industrial production. Harnessing the strengths of LMs is expected to unlock new opportunities and drive innovation across various industries.**

***Index Terms*—Large models, neural networks, smart industrial productions, artificial intelligence.**

I. Introduction

Large models (LMs) are neural networks with extremely large-scale parameters, enabling them to tackle tasks like content creation that were previously exclusive to humans. They represent a major shift in artificial intelligence (AI), marking its transition from a weak to a strong state. The development of AI has occurred in three distinct generations. The first generation, emerging in the 1970s, centered on computational intelligence, focusing on calculations and data storage. The second generation, which gained prominence in the 2000s, emphasized perceptive intelligence, enhancing recognition and interpretation across various modalities. By the 2020s, AI had progressed to cognitive intelligence, aimed at understanding and responding to external environments to support decision-making and task execution. LMs, as a key element of this third generation, are designed to analyze environmental inputs and make swift decisions, thereby assisting humans effectively.

Numerous groundbreaking discoveries have driven the evolution of LMs. The backpropagation algorithm [1] solved the challenges in the training of neural networks. The universal approximation theorem [2] demonstrated that neural networks with sufficient neurons can approximate any continuous function. The Transformer architecture [3] enhanced computational efficiency by removing loop structures and supporting parallel processing of long-term dependencies. Self-supervised learning methods [4] enabled effective training on unlabeled data. Additionally, the neural scaling law [5] revealed a predictable relationship between model performance, data volume, parameters, and computational power, showing that larger models and datasets improve results according to a power-law scale. Fundamentally, LMs rely on powerful algorithms and extensive computational resources to learn complex probability distributions from massive

This work was supported by the Beijing Natural Science Foundation under Grants L233005 and 4232049, the National Natural Science Foundation of China under Grants 62173013 and 62473014.

datasets, making their remarkable capabilities possible.

In recent years, LMs have experienced significant progress characterized by rapid growth in scale, enhanced capabilities, and broader application domains. Early LMs like GPT-1 laid the groundwork for more advanced systems, culminating in multimodal LMs such as GPT-4o. These developments have expanded the scope of natural language processing and built a solid foundation for future breakthroughs in AI. Despite these advancements, challenges such as excessive power consumption and catastrophic forgetting have become evident as LMs are integrated into real-world applications. Overcoming these obstacles is essential to further enhance LMs' capabilities and unlock their full potential in practical scenarios. This work examines the challenges associated with applying LMs and proposes potential solutions. It also explores the implementation of LMs in the industry.

II. Problems Encountered with LMs

A. Catastrophic Forgetting

Catastrophic forgetting denotes that training LMs on new tasks can lead to a decline in performance on previously learned tasks due to the model's inability to retain prior knowledge. This occurs because the model lacks an effective mechanism to preserve the data and scenarios encountered during earlier training phases. While fine-tuning a model with domain-specific data can improve its performance in a particular area, this often comes at the cost of general performance, especially when the new training data is significantly different from the original. Over time, the model may struggle to accurately interpret data that it initially trained on, resulting in poor performance. In healthcare applications, an LM trained in data related to a specific disease may perform well in diagnosing that condition. However, if the model is later fine-tuned with data on a different disease, it may lose its ability to effectively diagnose the original condition. This can reduce its reliability and accuracy, making it less effective in clinical applications where it needs to handle a broad range of medical scenarios.

B. Substantial Consumption of Resources

The number of parameters in a model directly affects the amount of data required for training. As LMs increase in size by reaching trillions of parameters, their demand for computational resources and large-scale datasets grows significantly. As illustrated in Fig. 1, AI's energy consumption continues to escalate. Both the quantity of training data and the computational power necessary for training scale proportionally with the number of model parameters [6]. This creates a substantial challenge in the training process, where expanding model capacity by increasing parameters leads to a corresponding rise in data and resource requirements. Additionally, by 2026, it is anticipated that the supply of high-quality text data may become insufficient, which would pose a major obstacle to further development of LMs.

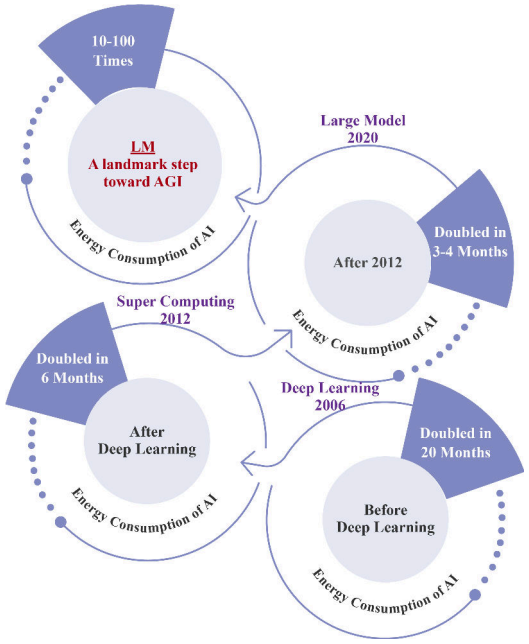

Fig. 1. Milestones in AI evolution and energy consumption trends.

Moreover, deep learning models require far more energy than human cognition. For instance, training a state-of-the-art image recognition model can consume 11,000 kWh of electricity [7], significantly higher than a human brain's energy expenditure during similar cognitive tasks. By comparison, the human brain operates on roughly 20 watts of power. This stark contrast in energy consumption underscores the high resource demands of modern AI systems, hindering the scalability and sustainability of LMs. As a result, researchers are focusing on more energy-efficient strategies such as model compression, knowledge distillation, and hardware acceleration to address these challenges.

C. Reasoning Deficits

The logical reasoning ability of LMs remains limited due to their black-box nature, which prevents them from employing structured problem-solving approaches. As

a result, they struggle with complex tasks that require advanced logical and numerical reasoning. The inverse scale phenomenon [8] suggests that increasing the number of parameters and training data may not always lead to improved performance, especially for tasks demanding higher-order cognitive functions. These limitations highlight that simply expanding the model does not address all challenges, particularly those requiring sophisticated reasoning. Moreover, the opacity of black-box models complicates the understanding of their decision-making processes and makes error correction difficult. To address these issues, researchers are exploring more interpretable machine learning approaches, such as rule-based systems, attention mechanisms, and explainable neural networks, which aim to provide clearer decision-making pathways and verifiable reasoning processes. These approaches are essential for enhancing the logical reasoning capabilities of LMs.

While LMs have limited logical reasoning abilities, they can still be applied at various stages of the optimization process [9]. Initially, LMs can assist in problem definition by transforming natural language descriptions into structured forms compatible with optimization algorithms. This helps clarify the problem, providing clear and actionable inputs for subsequent optimization processes. Furthermore, LMs can support the development of optimization strategies by analyzing relevant literature and technical documents, identifying suitable algorithmic approaches, and suggesting optimal parameter settings. This enables the extraction of key insights, guidelines, and best practices, improving the efficiency of optimization algorithms. Despite their reasoning limitations, these applications demonstrate the significant potential of LMs in addressing challenges in optimization.

D. Difficulty in Locating Errors

LMs often encounter difficulties in recognizing and correcting errors due to their inherent lack of self-awareness and limited capacity to understand the causes of mistakes. Therefore, this limitation impedes their ability to address and rectify errors effectively. For example, in processing tasks such as image captioning, LMs may produce inaccurate captions due to incorrect interpretations of visual data or biases in the training dataset. Thus, without mechanisms to reflect upon and analyze these errors, LMs cannot learn from them, hindering their continuous improvement. This underscores the need to develop more advanced error detection and correction strategies that facilitate the accurate identification of issues and prevent their recurrence. In addition, ensuring that LMs are trained on diverse and unbiased datasets is critical in mitigating errors at the outset. These limitations restrict their practical application, as the reliability and robustness of their outputs remain compromised. Researchers have focused on creating sophisticated feedback loops and error analysis frameworks to address these challenges. These efforts enable models to recognize errors, understand their underlying causes, and implement targeted adjustments. Furthermore, integrating human oversight with automated learning processes improves the model's capacity for self-correction and enhances its adaptability to new and complex situations. This human-machine collaborative approach is essential for improving the robustness and performance of LMs in real-world applications.

E. Underlying Issues Behind LM Limitations

1) Challenges of Back-Propagation in Training LMs: Back-propagation is the foundational mechanism used in training LMs, aiming to minimize global errors by iteratively adjusting model parameters. While this approach proves effective for optimizing performance on specific tasks, it requires significant computational resources and vast amounts of training data, making it susceptible to overfitting and reducing the model's overall flexibility [10]. Furthermore, when LMs are introduced to new tasks, they often need to start training from scratch, which is inefficient and limits their long-term scalability. This leads to a situation where task-specific parameters are optimized at the cost of others. When new tasks are incorporated, the back-propagation mechanism updates all parameters, causing the model to "forget" previously learned tasks. It underscores a major limitation of the current paradigm, as it struggles with continuous learning and adapting to new environments. Overcoming this challenge is critical and requires improvements in training and optimization strategies to enhance generalization and adaptability. These advancements would allow LMs to handle various tasks and maintain consistent performance over time.

During the reasoning, LMs generate answers based on the questions and parameters learned during training. The forward reasoning process, which involves all model layers and parameters, requires significant computational resources, leading to high energy consumption. A single ChatGPT interaction is estimated to consume over

a hundred times more energy than a Google search. This stark difference highlights the substantial energy overhead, which becomes a key concern in large-scale deployments and scenarios involving frequent invoke.

2) Challenges in the Architecture and Scalability of LMs: The architecture of current LMs is characterized by a combination of various modules and a hierarchical structure, which lacks clear functional definitions and does not integrate mechanisms in a way that aligns with human-understandable knowledge. This limitation restricts the model's ability to learn causal relationships, which are essential for effective reasoning, thereby impeding performance on more complex tasks. Furthermore, due to limited interoperability, LMs struggle to adapt to dynamic changes or new situations. Although LMs can handle large volumes of simultaneous text input, they face challenges when processing multimodal data from many requests concurrently, such as audio or video tokens. In such cases, the model may experience delays, queuing, or even system crashes, resulting in a suboptimal user experience. Consequently, there is a need for optimization in both the processing capacity and the architectural design of LMs to address these multimodal input challenges. Enhancing the stability and responsiveness of LMs under high-load conditions is also crucial to ensure their reliability and scalability in real-world applications.

III. Directions in Problem-Solving Efforts

A. Integrating Neural Networks and Symbolic Systems

Neural networks excel in leveraging abundant prior knowledge, demonstrating robust generalization and adaptability. However, they face significant challenges, including limited reasoning capabilities and low interpretability. In contrast, symbolic systems offer strengths such as composability, interpretability, and advanced reasoning, yet are hindered by issues like combinatorial explosion, sensitivity to noise, and limited generalization. To address these limitations, researchers are exploring hybrid paradigms that integrate the strengths of both systems in a complementary manner.

Potential strategies for integrating neural networks and symbolic systems include using symbolic frameworks with neural networks as supporting modules, designing neural networks to convert non-symbolic data into symbolic formats for further processing, and training neural networks with symbolic rule datasets to enhance reasoning capabilities. Additionally, symbolic rule-based structural templates can be embedded into neural architectures to guide their operations, while iterative feedback loops between neural networks and symbolic reasoning modules enable dynamic exchanges that enhance overall functionality. These integration approaches seek to leverage the complementary strengths of both paradigms, aiming to bridge the gap between the generalization power of neural networks and the interpretability and precision of symbolic reasoning. By combining these methodologies, LMs can better emulate human cognitive processes, allowing them to handle complex reasoning tasks with improved flexibility, accuracy, and decision-making quality, ultimately broadening their potential applications in diverse domains.

B. Optimizing Scalability and Decentralized Deployment

The advancement of AI fundamentally depends on computing power and generalized algorithms, as these elements far surpass the impact of individual skills in AI systems [11]. Kaplan *et al.* [12] highlight the critical role of the training scale, demonstrating that larger datasets and more expansive models consistently yield superior outcomes. The emergent properties [13] reveal that as models and datasets increase, LMs can unexpectedly acquire capabilities absent in smaller models. These emergent abilities suggest that the complexity of behavior and functionality in LMs is intricately linked to scale. To harness these properties, researchers focus on optimizing large-scale model designs and employing advanced optimization methods in the training process. These strategies aim to maximize the potential of large-scale systems while ensuring their reliability and effectiveness in practical applications.

Only scaling up models is no longer a practical solution, particularly in scenarios requiring simultaneous access by a large number of users. Centralized data centers often face significant challenges in efficiently handling such demands. Consequently, downsizing models and decentralizing computations to edge devices have emerged as critical strategies. Through model optimization and distributed computing, LMs can be effectively deployed on edge devices such as smartphones, IoT devices, and automotive systems. This approach enhances computational efficiency, strengthens data privacy, and reduces latency by processing data closer to its source.

However, edge devices are constrained by limited computational power, memory, storage, and energy efficiency, making the direct deployment of LMs challenging [14]. Addressing these limitations necessitates

advancements in model optimization techniques to improve performance while reducing resource consumption. Model compression plays a vital role in this context. Quantization lowers storage and computational demands by converting high-precision floating-point representations to low-precision formats. Pruning reduces model complexity by eliminating less significant parameters [15]. Knowledge distillation further contributes by transferring the capabilities of a larger LM to a smaller one, thereby significantly reducing the model size without compromising performance [16]. Furthermore, cloud-assisted mobile edge computing provides an effective complementary solution. In this paradigm, edge devices perform preliminary tasks such as preprocessing and feature extraction while computationally intensive operations are offloaded to the cloud. This hybrid approach reduces the computational burden on edge devices and ensures the scalable and efficient deployment of LMs in decentralized environments.

C. Referring to Human Memory Patterns

Brain science has heavily inspired neural network development, forming the underlying structure of contemporary LMs. Many challenges associated with LMs might be addressed by leveraging insights from how the human brain processes knowledge. Memory is the cornerstone of human intelligence, influencing cognitive functions such as learning, abstraction, reasoning, and association. These processes are shaped by three key stages: encoding, storage, and retrieval. In the human brain, encoding organizes and transforms external stimuli into meaningful representations, with learning efficiency closely tied to the strategies employed during this phase. Multi-modal encoding and context-based association are shown to enhance learning outcomes significantly. Storage involves categorizing and retaining knowledge in hierarchical long-term memory, enabling prior experience to facilitate future learning processes. Retrieval is the process of accessing stored information. It strengthens memory consolidation, promotes abstract thinking, and fosters reasoning by activating relevant associations. These mechanisms are foundational to human intelligence, as depicted in Fig. 2.

A critical difference between human cognition and current LMs lies in the mechanism of reasoning and memory utilization. In humans, reasoning leverages a selective retrieval process that activates only a small, relevant subset of long-term memory and converts it into working memory. This efficient mechanism contrasts sharply with LMs, which rely on simultaneously activating all parameters during reasoning, leading to significant inefficiencies. The human brain has approximately 10^{11} neurons and 10^{15} synapses. It operates on merely 20–23 watts of energy, while an LM of comparable scale may require up to 7.9×10^{6} watts, highlighting the vast energy disparity. Drawing inspiration from the human brain's retrieval and activation mechanisms offers potential pathways for enhancing machine intelligence. Developing models that mimic selective memory retrieval and adaptive activation could address LMs' heavy reliance on data and computational power, paving the way for more energy-efficient and intelligent systems.

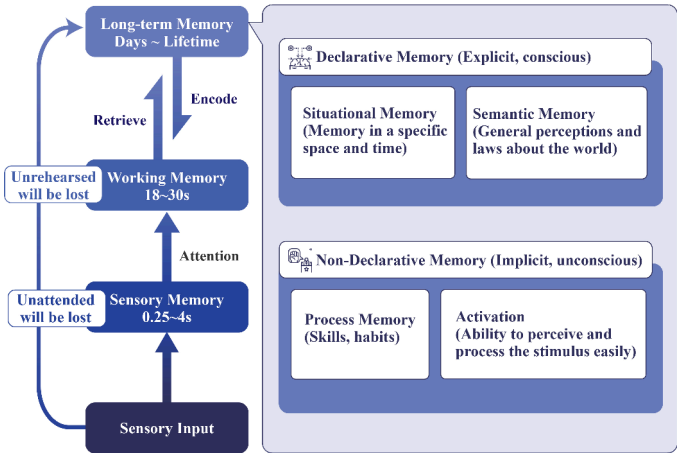

Fig. 2. Structure and Function of Human Memory.

IV. Large Model's Applications in Industry

LMs are used in intelligent industrial production, streamlining numerous tasks and boosting overall efficiency. These models facilitate industrial text generation and knowledge-based question-answering, enabling the automated creation of production handover reports, equipment inspection logs, and other essential documentation. Moreover, industrial multimodal LMs handle tasks such as helmet detection and defective product rate analysis, automating processes traditionally performed manually and significantly enhancing worker productivity. The architecture of industrial LMs is illustrated in Fig. 3. This framework processes various inputs, including structured and unstructured mission plans, natural language instructions, and multimodal environmental data. These inputs are directed to the model layer, where a suitable LM is chosen based on the task requirements.

To ensure reliability, the system may incorporate smaller models, rule-based mechanisms, or prebuilt knowledge bases to verify outputs and filter out erroneous instructions before execution. After task execution, the resulting state is reintroduced into the system as updated environmental data, enabling iterative improvement and responsiveness to dynamic industrial settings. This architecture is designed to cater to diverse industrial applications by integrating LMs into a feedback-driven loop. It optimizes routine processes and ensures the adaptability and safety of automated operations, making it a robust solution for modern manufacturing environments.

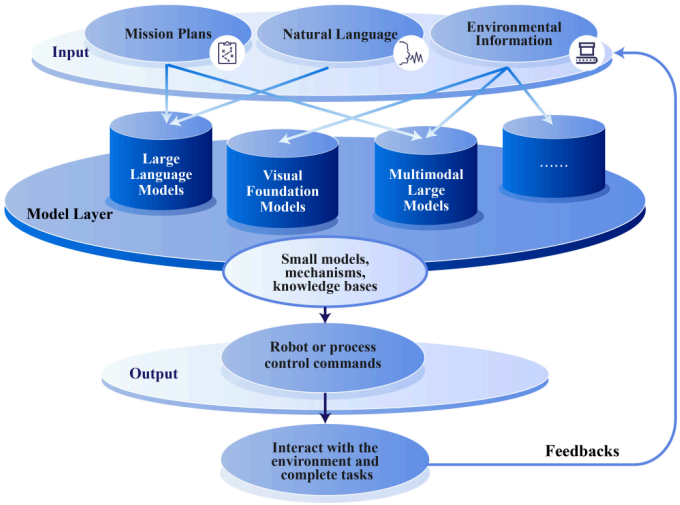

Fig. 3. Applications of Large Models in Industrial Scenarios.

V. Conclusions

The rapid advancement of Large Models (LMs) is transforming numerous industries by unlocking the potential of data for enhanced decision-making support. Despite their advantages, LMs still face challenges such as high data and energy requirements, catastrophic forgetting, and limited reasoning capabilities, which hinder their broader application. This work reviews the evolution of LMs, identifies their existing challenges, and explores potential solutions. Furthermore, it discusses the application of LMs in smart industrial systems. We believe that LMs will be pivotal in shaping the future and driving progress across various industries.

References

[1] X. Luo, H. Qu, Y. Wang, Z. Yi, J. Zhang and M. Zhang, "Supervised Learning in Multilayer Spiking Neural Networks With Spike Temporal Error Backpropagation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10141–10153, Dec. 2023.

[2] G. C. Calafiore, S. Gaubert and C. Possieri, "A Universal Approximation Result for Difference of Log-Sum-Exp Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5603–5612, Dec. 2020.

[3] Q. Zheng, Z. Peng, Z. Dang, L. Zhu, Z. Liu, Z. Zhang and J. Zhou, "Deep Tabular Data Modeling With Dual-Route Structure-Adaptive Graph Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9715–9727, Sept. 2023.

[4] J. Ruan, Q. Zheng, R. Zhao and B. Dong, "Biased Complementary-Label Learning Without True Labels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2616–2627, Feb. 2024.

[5] L. Zhang, S. Wang, J. Liu, X. Chang, Q. Lin, Y. Wu and Q. Zheng, "MuL-GRN: Multi-Level Graph Relation Network for Few-Shot Node Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 6085–6098, Jun. 2023.

[6] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. Rae and L. Sifre, "Training Compute-optimal Large Language Models," *36th International Conference on Neural Information Processing Systems*, 2022, New Orleans, USA, pp. 30016–30030.

[7] A. Luccioni, S. Viguier and A. Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," *Journal of Machine Learning Research*, no. 24, pp. 1–15, Jun. 2023.

[8] Y. Fu, C. Liu, D. Li, Z. Zhong, X. Sun, J. Zeng and Y. Yao, "Exploring Structural Sparsity of Deep Networks Via Inverse Scale Spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1749–1765, Feb. 2023.

[9] J. Bi, Z. Wang, H. Yuan, J. Zhang and M. Zhou, "Cost-Minimized Computation Offloading and User Association in Hybrid Cloud and Edge Computing," *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16672–16683, May 2024.

[10] H. Shao, Y. Zou, C. Liu, Q. Guo and D. Zhong, "Learning to Generalize Unseen Dataset for Cross-Dataset Palmprint Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3788–3799, May 2024.

[11] Sutton R. The Bitter Lesson. 2019.

[12] J. Kaplan, S. McCandlish, T. Henighan, T. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei, "Scaling Laws for Neural Language Models," *arXiv*, https://doi.org/10.48550/arXiv.2001.08361.

[13] G. -J. Qi and J. Luo, "Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2168–2187, Apr. 2022.

[14] H. Yuan, J. Bi, Z. Wang, J. Yang and Jia Zhang, "Partial and Cost-minimized Computation Offloading in Hybrid Edge and Cloud Systems," *Expert Systems with Applications*, vol. 250, pp. 1–13, Sept. 2024.

[15] S. Fu, F. Dong, D. Shen and T. Lu, "Privacy-preserving Model Splitting and Quality-aware Device Association for Federated Edge Learning," *Software: Practice and Experience*. vol. 54, no. 10, pp. 2063—2085, May 2024.

[16] R. Yang and C. Deng, "Surpass Teacher: Enlightenment Structured Knowledge Distillation of Transformer," *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Honolulu, Oahu, HI, USA, 2023, pp. 5102–5107.