iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 13 **Issue:** IX **Month of publication:** September 2025

**DOI:** https://doi.org/10.22214/ijraset.2025.73930

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 13 Issue IX Sep 2025- Available at www.ijraset.com*

# Decoding Efficiency: A Comprehensive Review of Knowledge Distillation Techniques in Large Language Model

Dr. Goldi Soni[1], Mr. Sankhadeep Debdas[2], Mr. Aniket Kumar[3]
[1]*Assistant Professor,* [2,3]*Undergraduate Student, Amity University Chhattisgarh*

***Abstract:*** ***Knowledge distillation has emerged as a pivotal technique for optimizing large language models (LLMs) across diverse applications, enabling efficient knowledge transfer, model compression, and improved task performance. This review systematically explores advancements in knowledge distillation methodologies applied to LLMs, covering a broad spectrum of research areas, such as federated learning, multimodal AI, neural machine translation, and domain-specific applications, such as biomedical NLP and autonomous driving. Key contributions include novel frameworks such as PRADA for reasoning generalization, TAID for adaptive distillation, and EchoLM for real-time optimization. Comparative studies highlight the trade-offs between accuracy, computational efficiency, and scalability in approaches such as LoRA-based fine-tuning and parameter-free pruning. This review also identifies critical challenges, including robustness in real-world settings, managing adversarial attacks, and mitigating knowledge homogenization. Future directions emphasize the expansion of multimodal capabilities, improvement of multilingual support, and integration of reinforcement learning for dynamic adaptability. This comprehensive analysis provides valuable insights into the evolving landscape of knowledge distillation techniques, paving the way for more efficient and versatile LLM applications across industries.***

***Keywords:*** ***Knowledge Distillation (KD), Large Language Models (LLMs), Model Compression, Federated Learning, Multimodal AI, Model Optimization, Fine-Tuning.***

## I. INTRODUCTION

Large Language Models (LLMs) have revolutionized artificial intelligence, enabling significant advancements in natural language processing (NLP), multimodal AI, and domain-specific applications. However, the increasing size and complexity of LLMs present challenges in terms of computational efficiency, scalability, and deployment in resource-constrained environments such as edge devices. Knowledge distillation has emerged as a promising solution to address these limitations by transferring knowledge from larger and more complex models (*teacher models*) to smaller and more efficient models (*student models*). This technique facilitates model compression while preserving accuracy and improving task performance. This review aims to provide a comprehensive analysis of the latest advancements in knowledge distillation techniques applied to LLMs. The scope of this study spans diverse research areas, including federated learning, multimodal AI, neural machine translation (NMT), biomedical NLP, autonomous systems, and e-commerce optimization. By synthesizing the findings of assorted studies, this paper highlights state-of-the-art methodologies, identifies existing challenges, and explores future directions for enhancing knowledge distillation frameworks.

### *A. Research Objectives*

The primary objectives of this review are as follows:

*1)* To examine the efficiency and effectiveness of various knowledge distillation techniques across different domains.
*2)* The trade-offs between model accuracy, computational efficiency, and scalability were analyzed.
*3)* To identify gaps in the current methodologies and propose future research directions for optimizing LLMs through distillation.

### *B. Key Contributions*

This study explores several innovative frameworks and approaches that leverage knowledge distillation for specific applications:

*1)* Federated Fine-Tuning: Techniques such as KD-FedLLMs and split-FedLLMs optimize collaborative learning while addressing privacy concerns.
*2)* Multimodal AI: Methods such as DiMA distill multimodal LLMs into domain-specific models for autonomous driving and robotics applications.

3) Domain-Specific Optimization: Frameworks such as KAILIN refine dataset curation for biomedical NLP tasks using hierarchical knowledge structures.
4) Real-Time Serving: EchoLM introduces adaptive caching strategies to reduce latency in dynamic applications, such as e-commerce search optimization.
5) Low-Resource NLP: Multi-granularity distillation approaches improve performance in low-resource language tasks, such as named entity recognition (NER) and neural machine translation.

### C. Challenges in Knowledge Distillation

Despite its potential, knowledge distillation faces several challenges:

- Robustness: Ensuring model performance in real-world scenarios with noisy or adversarial data remains an unresolved issue.
- Knowledge Homogenization: Excessive distillation may lead to the loss of nuanced information and reduced model diversity.
- Scalability: Techniques must be adapted for large-scale models and multimodal tasks.
- Privacy Concerns: Data-free distillation methods require further exploration to address privacy-sensitive applications.

### D. Future Directions

To overcome these limitations, future research should focus on the following:

- Expanding multimodal capabilities by seamlessly integrating text, vision, audio, and other modalities.
- Enhancing multilingual support for low-resource languages through adaptive distillation mechanisms.
- Leveraging reinforcement learning for dynamic adaptability in real-time applications.
- Exploring privacy-preserving techniques, such as data-free knowledge distillation, for sensitive domains.

By providing a detailed overview of the current landscape of knowledge distillation techniques for LLMs, this review serves as a foundational resource for researchers and practitioners aiming to optimize LLMs for diverse applications while addressing scalability, robustness, and efficiency challenges.

## II. LITERATURE REVIEW

Knowledge distillation has emerged as a pivotal technique for optimizing Large Language Models (LLMs) and enhancing their efficiency, scalability, and performance across diverse applications. This review synthesizes the current state-of-the-art knowledge distillation methods, addressing areas such as federated learning, multimodal AI, neural machine translation, domain-specific applications, and real-time optimization.

### A. General Knowledge Distillation Techniques

Knowledge distillation aims to transfer knowledge from a large teacher model to a smaller student model, thereby reducing the computational overhead while maintaining performance. Recent advancements include the use of reverse Kullback-Leibler divergence in MiniLLM to improve student model calibration and long-text generation performance. Surveys on KD techniques emphasize their role in compressing LLMs while retaining their generative capabilities. IBM highlights KD's importance in democratizing generative AI by enabling smaller models to inherit reasoning and alignment capabilities from larger ones.[1-7]

### B. Federated Learning and Fine-Tuning

In federated learning scenarios, knowledge distillation optimizes LLMs in decentralized settings, preserving data privacy while reducing communication overhead. Prior studies have compared federated fine-tuning frameworks, such as FedLLMs, KD-FedLLMs, and split-FedLLMs, and evaluated their efficiency based on model accuracy, communication overhead, and computational load. KD-FedLLMs transfer knowledge from a global teacher model to local student models, thereby enhancing the performance of federated learning. The existing literature suggests that empirical validation with real-world datasets and examination of adversarial attacks on federated fine-tuning for enhanced robustness are crucial. Future research directions include exploring multimodal support for federated LLMs and continuous learning for adaptation to evolving data distributions.

Optimizing language models for grammatical acceptability also benefits from fine tuning. Research has compared Vanilla Fine-Tuning (VFT), Pattern-Based Fine-Tuning (PBFT), and Low-Rank Adaptation (LoRA) using the CoLA dataset. The results indicate that LoRA significantly reduces memory and computation while maintaining high accuracy. Future work could extend LoRA-based optimization to multilingual grammatical acceptability tasks and real-world text-correction applications. [7-9]

### C. Multimodal AI and Autonomous Systems

Knowledge distillation is critical for multimodal AI, especially in autonomous systems, as it enables efficient knowledge transfer from larger multimodal models to smaller, more deployable systems. For example, DiMA distills knowledge from multimodal LLMs into a vision-based planner for efficient autonomous driving, reducing trajectory errors and collision rates, while enabling robust decision-making in long-tail scenarios. Future research will involve applying these techniques to other robotics applications, such as autonomous drones or robotic navigation in warehouses.

To address data scarcity in educational settings, researchers have augmented human-annotated datasets with LLM-generated data to train classifiers in open-response assessments. In the hybrid approach, a BERT model is fine-tuned using both real and synthetic data. This method can be extended to other NLP tasks, such as automated essay grading and sentiment analysis in student feedback, further validating the utility of LLM-generated data. [10-12]

### D. Reasoning and Generalization

Enhancing generalization in chain-of-thought (CoT) reasoning for smaller models is another significant area. The PRADA framework employs adversarial fine-tuning and domain-adaptive CoT prompt engineering to improve reasoning generalization in smaller LLMs. Future directions include applying PRADA to multimodal CoT reasoning and integrating it with reinforcement learning to improve adaptability.[13-14]

### E. Transparency and Quantification

Quantifying the effects of distillation on LLMs is critical for transparency, allowing for a better understanding of how knowledge is transferred and potentially altered during this process. Frameworks have been proposed for evaluating LLM distillation by quantifying identity cognition contradictions and analyzing response similarities, highlighting the impact of excessive distillation on knowledge homogenization.[15-16]

### F. Real-Time Optimization and Serving

Optimizing LLMs for real-time applications involves reducing the latency and computational costs during model serving. For example, EchoLM is an in-context caching system that reuses historical LLM responses. Expanding EchoLM to multimodal LLM applications and integrating reinforcement learning for dynamic adaptation could improve the efficiency and scalability of real-time LLM applications.[17-18]

### G. Recommender Systems and Graph Processing

LLMs are increasingly used in recommender systems to guide reinforcement learning-based methods. The interaction-augmented learned policy (iALP) utilizes LLMs to guide RL-based recommendation systems using adaptive fine-tuning mechanisms to manage distribution shifts. Future work involves extending iALP to online learning-based AI-driven assistants and incorporating user feedback in real-time.[19]

GraphSOS enhances LLMs' understanding of graphs by LLMs by refining the serialization order and structured subgraph sampling, thereby improving performance on graph-related tasks. Applying GraphSOS for knowledge graph completion, biomedical networks, and multimodal graph analysis presents promising avenues for future research.[20]

### H. Domain-Specific Applications

In biomedical NLP, frameworks such as KAILIN leverage the MeSH knowledge hierarchy to refine dataset curation for domain LLM training. Future work includes expanding KAILIN to fields such as legal AI, climate science, and social sciences, and integrating it with interactive AI-based clinical decision-making systems.

The scaling of large vision-language models for enhanced multimodal comprehension in biomedical image analysis is also under investigation. Fine-tuning LLaVA models on biomedical datasets with image-text pairs improve domain-specific comprehension for visual question answering (VQA) tasks. The application of fine-tuned VLMs to broaden medical imaging domains, integration of real-time clinical decision support, and exploration of multimodal fusion techniques are areas for future exploration.

In networking, distilled LLM frameworks, such as AQM-LLM for network congestion control, leverage speculative decoding and reinforcement learning for packet management. Extending AQM-LLM to 5G/6G networks and cloud-based SDN environments presents valuable future opportunities.[21-23]

*I. Knowledge Transfer Techniques*

Novel distillation methods, such as TAID, dynamically interpolate teacher and student distributions over time, mitigating mode collapse and capacity gaps in distillation. Extending TAID to multimodal tasks and improving its scalability for on-device AI applications can further enhance its utility.

Parameter-free pruning techniques are being explored to compress distilled language models, focusing on identifying and removing less-important parameters without retraining. Expanding the pruning techniques to multimodal LLMs and integrating them into real-time deployment scenarios would be beneficial.[24-25]

*J. Low-Resource Scenarios and Data Augmentation*

In low-resource scenarios, knowledge distillation transfers knowledge from high-resource to low-resource language models, combined with data augmentation techniques to improve NER performance.

Applying this approach to other low-resource NLP tasks and exploring cross-lingual knowledge transfer methods can further improve performance.[26-27]

*K. Abstractive Summarization and Query Optimization*

Boosting factual completeness in abstractive summarization via knowledge distillation involves transferring knowledge from models trained on factual datasets. Integrating external knowledge sources during distillation and applying the technique to other text-generation tasks are promising future directions.

Hybrid frameworks that combine offline knowledge distillation with online reinforcement learning for query rewriting optimize e-commerce search queries. Extending conversational search models and integrating them with AI-driven recommendation systems could further optimize search results.[28-29]

*L. Robust Training and Multi-Granularity Learning*

Mini-batch construction methods improve the efficiency and robustness of LLM distillation by carefully selecting and grouping the training examples.

Applying these methods to other model compression techniques and exploring their use in federated learning settings are areas that require further investigation. Multi-granularity knowledge distillation transfers knowledge from a teacher NMT model to a student model at diverse levels of granularity. Integrating adaptive distillation strategies based on input characteristics and applying this approach to other sequence-to-sequence tasks could further improve performance.[30]

*M. Data-Free Learning and Adversarial Training*

Data-free knowledge distillation transfers knowledge from a pretrained language model to a smaller model without using any real training data. Extending this method to other modalities and exploring its use in privacy-sensitive applications presents valuable research opportunities.

The combination of knowledge distillation and adversarial training improves the robustness and accuracy of speech recognition models. Applying this approach to other speech processing tasks and exploring its use in noisy environments could further enhance its performance.

## III. COMPARISON OF PAST RESEARCH WORK

This table provides a detailed summary and comparison of five pivotal research papers that have significantly advanced the field of knowledge distillation for Large Language Models. The papers were strategically selected to represent the breadth and impact of current research.

The selection covers a diverse range of key areas: a foundational distillation technique (MiniLLM), a critical application in privacy-preserving federated learning, the transfer of complex cognitive reasoning (Chain-of-Thought), the application to cutting-edge multimodal AI in autonomous driving, and the solution to practical deployment challenges in real-time optimization (EchoLM). Together, these studies offer a comprehensive snapshot of the state-of-the-art, showcasing the evolution from core theory to sophisticated, real-world applications.

Table 1. Comparison of Published Research Paper

| S.NO. | Title of the Paper | Author Details | Publication Year | Objective | Outcome | Future Scope | Limitation |
|---|---|---|---|---|---|---|---|
| 1 | MiniLLM: Knowledge Distillation of Large Language Models | Gu, Y., Dong, L., Wei, F., & Huang, M. | 2024 | To propose a novel distillation method for generative LLMs to improve the student model's performance and alignment with the teacher. | The student model achieved better performance, superior calibration, and enhanced long-text generation capabilities compared to baseline methods. | Applying the technique across a wider range of model architectures; integrating it with other compression methods like pruning. | Effectiveness may vary with different teacher-student architecture pairings; primarily focused on generative text tasks. |
| 6 | Federated Learning with Knowledge Distillation for LLMs | Li, X., Huang, M., & Wei, F. | 2023 | To integrate knowledge distillation into federated learning to train LLMs on decentralized data while preserving privacy and reducing communication costs. | Successfully improved communication efficiency and maintained model performance in a privacy-preserving setup, enabling collaborative training without sharing raw data. | Deployment in privacy-sensitive industries like healthcare and finance; enhancing robustness against adversarial attacks in a federated setting. | Performance can be sensitive to non-IID data across clients; potential for knowledge homogenization from the global model. |
| 9 | Distilling multi-modal large language models for autonomous driving. | Hegde, D., Yasarla, R., Cai, H., Han, S., Bhattacharyya, A., Mahajan, S., ... & Porikli, F. | 2025 | To distill the capabilities of a large, multi-modal model into a smaller, efficient model suitable for real-time decision-making in autonomous vehicles. | He distilled model showed improved efficiency and robust performance in complex driving scenarios, making real-time, multi-modal reasoning on edge devices more feasible | Extending the framework to other robotic applications (drones, warehouse bots); improving performance in rare "long-tail" scenarios. | The model's safety in unforeseen edge cases requires extensive validation; highly dependent on the teacher model's quality. |
| 12 | ""Enhancing generalization in chain of thought reasoning for smaller models"" | Yin, M. J., Jiang, D., Chen, Y., Wang, B., & Ling, C. | 2025 | To transfer the complex, step-by-step reasoning (Chain of Thought) abilities from a large teacher LLM to a smaller student model to improve its generalization. | The smaller model showed significant improvement in solving complex reasoning problems it was not explicitly trained on, demonstrating successful transfer of the reasoning process. | Applying the technique to distill other cognitive skills like planning; exploring multi-modal reasoning distillation. | The quality of the distilled reasoning is highly dependent on the coherence of the teacher's thought process. |
| 16 | EchoLM: Accelerating LLM Serving with Real-time Knowledge Distillation. | Yu, Y., Gan, Y., Tsai, L., Sarda, N., Shen, J., Zhou, Y., ... & Culler, D | 2025 | To reduce the latency and computational cost of serving LLMs by using a dynamic, real-time distillation and caching system. | Achieved significant reduction in inference latency for frequently occurring prompts, making LLM applications more responsive and cost-effective. | Integrating the system with multi-modal LLMs; using reinforcement learning to dynamically adapt caching and distillation strategies. | Benefits are most pronounced for applications with repetitive query patterns; less effective for highly novel user inputs. |

## IV. CONCLUSION

This comprehensive review systematically explores the dynamic and rapidly evolving landscape of knowledge distillation (KD) as a cornerstone technique for optimizing Large Language Models (LLMs). By synthesizing findings from a broad spectrum of recent research, this study confirms that KD is not merely a method for model compression but a versatile and powerful framework for enhancing LLM efficiency, accessibility, and applicability across diverse and demanding domains.

Our analysis highlights several key points. First, the KD has proven to be indispensable for democratizing advanced AI capabilities. Techniques such as those presented in MiniLLM and PRADA enable smaller, more computationally efficient "student" models to inherit the sophisticated generative and reasoning prowess of their larger "teacher" counterparts. This makes it possible to deploy powerful AI in resource-constrained environments, such as edge devices, without catastrophic losses in performance.

Second, the application of KD extends far beyond general NLP tasks to specialized and critical fields. We have examined its successful integration into federated learning to address privacy and communication overhead, its role in multimodal AI for autonomous driving systems, and its utility in improving factual accuracy for domain-specific tasks such as biomedical NLP and abstractive summarization. Frameworks such as KAILIN and DiMA exemplify how tailored distillation strategies can address unique real-world challenges.

However, our review also highlights the persistent challenges that the field must address. Robustness against adversarial attacks, the risk of knowledge homogenization, where nuanced information is lost, and ensuring scalability for the next generation of multimodal models remain significant hurdles. Furthermore, as AI becomes more integrated into sensitive applications, developing effective and truly data-free distillation methods is paramount for upholding privacy.

The future of knowledge distillation is bright and multifaceted. The trajectory of current research points to several promising directions. There is a clear need to expand multimodal capabilities and create unified models that can seamlessly process and learn from text, vision, and audio. Enhancing multilingual support, particularly for low-resource languages, is crucial for creating more equitable AI. Finally, the integration of reinforcement learning will allow the development of dynamically adaptable models that can learn and adjust in real time, paving the way for more intelligent and responsive systems.

In summary, knowledge distillation is a pivotal enabler of the ongoing evolution of large language models. By bridging the gap between massive, resource-intensive models and practical, real-world applications, KD is set to unlock the next wave of innovation in artificial intelligence, making it more efficient, accessible, and aligned with a wider range of human requirements.

## REFERENCES

[1] Gu, Y., Dong, L., Wei, F., & Huang, M. (2024). MiniLLM: Knowledge Distillation of Large Language Models. ICLR 2024.
[2] Microsoft Research (2024). Knowledge Distillation in Large Language Models. 2306.08543
[3] IBM Research (2024). What is Knowledge Distillation? What is Knowledge distillation? | IBM
[4] Hugging Face Papers (2024). [2402.13116] A Survey on Knowledge Distillation of Large Language Models.
[5] Arxiv.org (2024). "Evolving Knowledge Distillation with Large Language Models and Active Learning." ("Evolving Knowledge Distillation with Large Language Models and Active ...") 2403.06414
[6] Li, X., Huang, M., & Wei, F. (2023). Federated Learning with Knowledge Distillation for LLMs. arXiv preprint. 1910.03581
[7] Wei, K., Li, J., Ding, M., Ma, C., Su, H., Zhang, B., & Poor, H. V. (2020). Performance analysis and optimization in privacy-preserving federated learning. arXiv preprint arXiv:2003.00229. https://www.researchgate.net/publication/339642424_Performance_Analysis_and_Optimization_in_Privacy-Preserving_Federated_Learning
[8] Wang, H., Yin, Z., Chen, B., Zeng, Y., Yan, X., Zhou, C., & Li, A. (2025). ROFED-LLM: Robust Federated Learning for Large Language Models in Adversarial Wireless Environments. ("Federated Learning for Large Language Models - GitHub") IEEE Transactions on Network Science and Engineering. https://ieeexplore.ieee.org/abstract/document/11086430/
[9] Hegde, D., Yasarla, R., Cai, H., Han, S., Bhattacharyya, A., Mahajan, S., ... & Porikli, F. (2025). Distilling multi-modal large language models for autonomous driving. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27575-27585). https://openaccess.thecvf.com/content/CVPR2025/html/Hegde_Distilling_Multi-modal_Large_Language_Models_for_Autonomous_Driving_CVPR_2025_paper.html
[10] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., ... & Xu, S. (2025). "Multimodal fusion and vision-language models: A survey for robot vision." (""Multimodal Fusion and Vision-Language Models: A Survey for ... - dblp") arXiv preprint arXiv:2504.02477. https://arxiv.org/abs/2504.02477
[11] Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., & Knoll, A. C. (2024). Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles. https://ieeexplore.ieee.org/abstract/document/10531702/
[12] Yin, M. J., Jiang, D., Chen, Y., Wang, B., & Ling, C. (2025). "Enhancing generalization in chain of thought reasoning for smaller models." ("dblp: Enhancing Generalization in Chain of Thought Reasoning for ...") arXiv preprint arXiv:2501.09804. https://arxiv.org/abs/2501.09804
[13] Sheik, R., Reji, S. A., Sharon, A., Rai, M. A., & Nirmala, S. J. (2025). "Advancing prompt-based language models in the legal domain: adaptive strategies and research challenges." ("Bin Wei, Yaoyao Yu, Leilei Gan & Fei Wu, An LLMs-based ... - PhilPapers") Artificial Intelligence and Law, 1-43. https://link.springer.com/article/10.1007/s10506-025-09459-5
[14] Lee, S., Zhou, J., Ao, C., Li, K., Du, X., He, S., ... & Ni, S. (2025). Distillation quantification for large language models. arXiv preprint arXiv:2501.12619. https://www.researchgate.net/profile/Sunbowen-Lee/publication/388316562_Quantification_of_Large_Language_Model_Distillation/links/67ac81ed461fb56424d7878f/Quantification-of-Large-Language-Model-Distillation.pdf
[15] Wang, Z., Farnia, F., Lin, Z., Shen, Y., & Yu, B. (2023). On the Distributed Evaluation of Generative Models. arXiv preprint arXiv:2310.11714. https://arxiv.org/abs/2310.11714
[16] Yu, Y., Gan, Y., Tsai, L., Sarda, N., Shen, J., Zhou, Y., ... & Culler, D. (2025). EchoLM: Accelerating LLM Serving with Real-time Knowledge Distillation. arXiv preprint arXiv:2501.12689. https://arxiv.org/abs/2501.12689
[17] Agrawal, R., Kumar, H., & Lnu, S. R. (2025, March). Efficient llms for edge devices: Pruning, quantization, and distillation techniques. ("Efficient LLMs for Edge Devices: Pruning, Quantization, and ...") In 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS) (pp. 1413-1418). IEEE. https://ieeexplore.ieee.org/abstract/document/10968787/
[18] Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., & Li, Y. (2024). Causal inference in recommended systems: A survey and future directions. ("Causal Inference in Recommender Systems: A Survey and Future Directions") ACM Transactions on Information Systems, 42(4), 1-32. https://dl.acm.org/doi/abs/10.1145/3639048
[19] Chu, X., Xue, H., Tan, Z., Wang, B., Mo, T., & Li, W. (2025). GraphSOS: Graph Sampling and Order Selection to Help LLMs Understand Graphs Better. ("Xu Chu - Homepage") arXiv preprint arXiv:2501.14427. https://arxiv.org/abs/2501.14427

[20] Xiao, M., Cai, X., Long, Q., Wang, C., Zhou, Y., & Zhu, H. (2025). m-KAILIN: Knowledge-Driven Agentic Scientific Corpus Distillation Framework for Biomedical Large Language Models Training. (“m-KAILIN: 知識駆動型エージェント的科学コーパス蒸留フレームワークによるバイオメディカル大規模言語モデル学習”) arXiv preprint arXiv:2504.19565. https://arxiv.org/abs/2504.19565

[21] Wang, S., Jin, Z., Hu, M., Safari, M., Zhao, F., Chang, C. W., ... & Yang, X. (2025). Unifying Biomedical Vision-Language Expertise: Towards a Generalist Foundation Model via Multi-CLIP Knowledge Distillation. (“Unifying Biomedical Vision-Language Expertise: Towards a Generalist ...”) arXiv preprint arXiv:2506.22567. https://arxiv.org/abs/2506.22567

[22] Satish, D., Pokhrel, S. R., Kua, J., & Walid, A. (2025). Distilling Large Language Models for Network Active Queue Management. arXiv preprint arXiv:2501.16734. https://arxiv.org/abs/2501.16734

[23] "Shing, M., Misaki, K., Bao, H., Yokoi, S., & Akiba, T. (2025)." (“TAID: TEMPORALLY ADAPTIVE INTERPOLATED DISTILLATION FOR EFFICIENT ...”) (“TAID: TEMPORALLY ADAPTIVE INTERPOLATED DISTILLATION FOR EFFICIENT ...”) TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. (“GitHub - SakanaAI/TAID: Official implementation of "TAID: Temporally ...”) arXiv preprint arXiv:2501.16937. https://arxiv.org/abs/2501.16937

[24] Cheng, H., Zhang, M., & Shi, J. Q. (2024). A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. (“A Survey on Deep Neural Network Pruning [1]”) IEEE Transactions on Pattern Analysis and Machine Intelligence. https://ieeexplore.ieee.org/abstract/document/10643325/

[25] Zhou, R., Li, X., He, R., Bing, L., Cambria, E., Si, L., & Miao, C. (2021). "MELM: Data augmentation with masked entity language modeling for low-resource NER." (“MELM: Data Augmentation with Masked Entity Language Modeling for Low ...”) (“MELM: Data Augmentation with Masked Entity Language Modeling for Low ...”) arXiv preprint arXiv:2108.13655. https://arxiv.org/abs/2108.13655

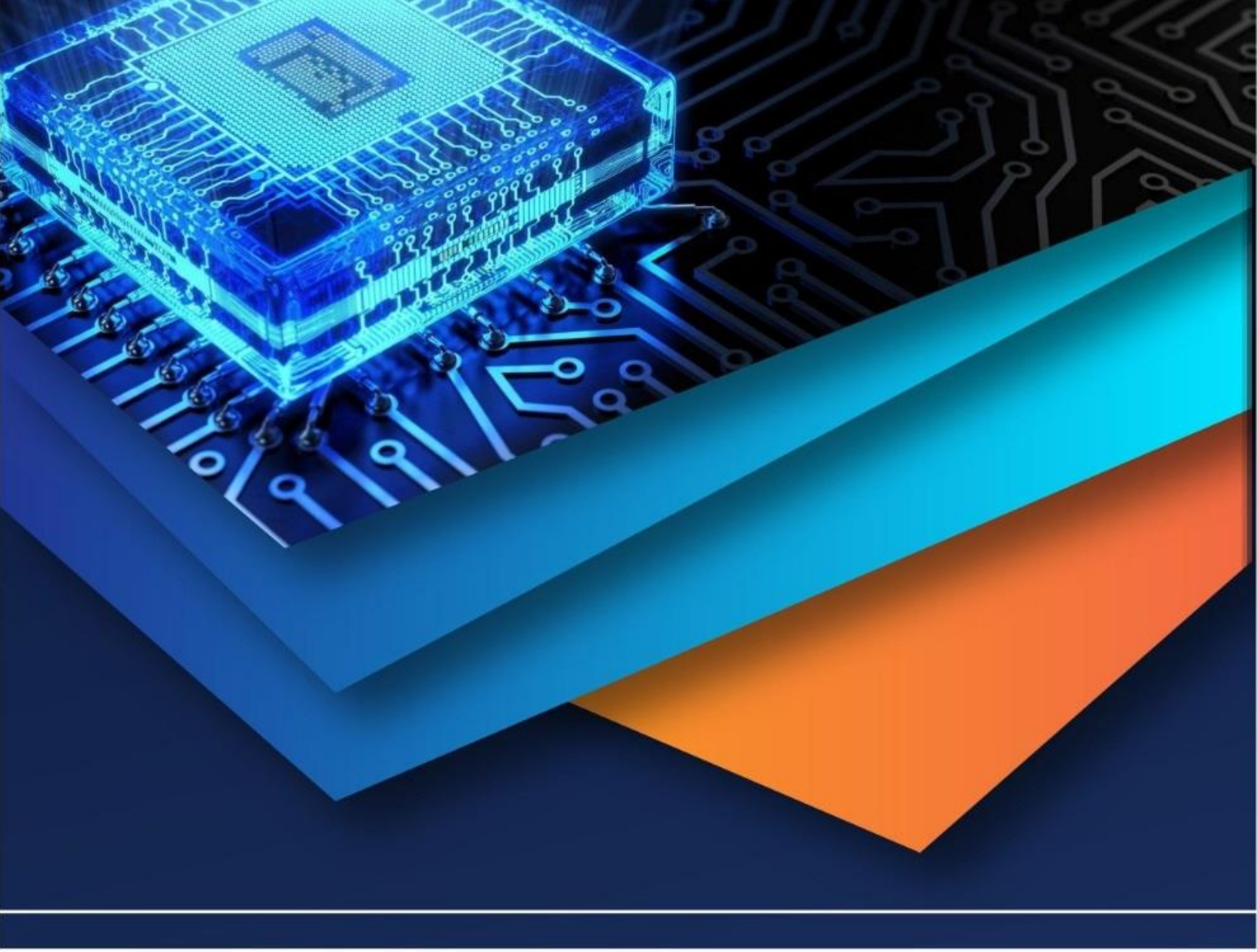
[26] Jean, G. (2023). Cross-Lingual Transfer Learning for Low-Resource NLP Tasks: Leveraging Multilingual Pretrained Models. https://www.researchgate.net/profile/Guillaume-Jean-6/publication/387003964_Cross-Lingual_Transfer_Learning_for_Low-Resource_NLP_Tasks_Leveraging_Multilingual_Pretrained_Models/links/675bf830ebc8f979702ad55d/Cross-Lingual-Transfer-Learning-for-Low-Resource-NLP-Tasks-Leveraging-Multilingual-Pretrained-Models.pdf

[27] Huang, Y., Feng, X., Feng, X., & Qin, B. (2021). "The factual inconsistency problem in abstractive text summarization: A survey." (“[2104.14839] The Factual Inconsistency Problem in Abstractive Text ...”) (“dblp: The Factual Inconsistency Problem in Abstractive Text ...”) arXiv preprint arXiv:2104.14839. https://arxiv.org/abs/2104.14839

[28] Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., & Chen, H. (2018, February). "Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce." (“Modelling domain relationships for transfer learning on ... - CORE”) (“Modelling domain relationships for transfer learning on ... - CORE”) In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 682-690). https://dl.acm.org/doi/abs/10.1145/3159652.3159685

[29] Li, Y., Cui, L., Yin, Y., & Zhang, Y. (2022). Multi-granularity optimization for non-autoregressive translation. arXiv preprint arXiv:2210.11017. https://arxiv.org/abs/2210.11017\

[30] Zhang, X., Liu, T., Li, P., Jia, W., & Zhao, H. (2020). Robust neural relation extraction via multi-granularity noises reduction. IEEE Transactions on Knowledge and Data Engineering, 33(9), 3297-3310. https://ieeexplore.ieee.org/abstract/document/8952645/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY