Faster Than the Teacher, Smarter Than the Student: Classifying with Wisdom via Knowledge Distillation in LLMs

Gobihanath B.[1], Abishethvarman V.[1]*, Prasanth S.[2], Banujan K.[3] and B.T.G.S Kumara [1,4]
[1]Department of Computing Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka
[2]Faculty of Engineering Applied Science, Memorial University of Newfoundland, Canada
[3]Faculty of Science and Engineering, Southern Cross University, Gold Coast, 4225, Australia
[4]Department of Software Engineering, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka
*Corresponding author: abishethvarman@gmail.com

ABSTRACT

Large language models (LLMs) have achieved remarkable success across various natural language processing (NLP) tasks, driven by their ability to capture complex language patterns through large-scale pretraining. However, their substantial computational demands limit their deployment in resource-constrained environments. To address this, this research introduced Knowledge distillation-based framework for text classification using a multiclass approach across three domains: entertainment, sports, and politics. We utilize both hard labels (ground-truth categories) and soft labels (logits from a teacher model) to train a student and a distilled model. The teacher model is accurate but computationally expensive. The student model is lightweight and fast, yet less accurate. Through distillation, we derive a task-specific distilled model that balances speed and accuracy. We also compare the performance of these models against traditional classifiers such as LSTM, SVM, and Naive Bayes. Traditional models excel comparing to the LLMs. Considering only task agnostic language models, evaluation shows that the distilled model performs significantly better than the student and competitively against the teacher, offering a practical trade-off. Our study demonstrates the value of soft label transfer and semantic alignment for improving classification performance in resource-constrained environments. The text classification code can be found at: https://github.com/Abishethvarman/KD-Text-Classification

KEYWORDS: *Knowledge Distillation, Text Classification, Multiclass Classification, Teacher-Student Framework.*

INTRODUCTION

The advent of the Transformer architecture has (Vaswani et al., 2017) revolutionized the field of natural language processing (NLP), enabling the development of powerful large language models (LLMs). Transformers rely entirely on attention mechanisms to model long-range dependencies in text, replacing traditional recurrent and convolutional structures. Building on this foundation, models such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), T5, and PaLM (Chowdhery et al., 2023) demonstrated unprecedented performance across a wide range of NLP tasks through pre-training on large text corpora and fine-tuning or prompting techniques. More recent advances have led to the emergence of multimodal and highly scalable models such as GPT-4, Gemini, LLaMA 2, and Gemma, capable of handling multiple input modalities and processing extensive contexts. These LLMs exemplify the scalability and generalization potential of the Transformer architecture, enabling few-shot, zero-shot, and even instruction-following abilities with minimal task-specific supervision.

Text classification (Kowsari et al., 2019), is a fundamental task in natural language processing, vital for applications like news categorization, sentiment analysis (Miah et al., 2024), and content filtering (Vörös et al., 2023). The introduction of large language models significantly changed the landscape of Natural Language Processing (NLP), providing new capabilities across a wide range of tasks like translation, summarization, classification and question answering. While these large language models (LLMs) offer remarkable performance in classification tasks, they are often computationally intensive. This makes the

https://doi.org/10.54389/LQAK9453

implementation difficult in situations with limited hardware capacity, such as mobile applications, real-time systems, and edge computing devices. Furthermore, the operational costs associated with proprietary models may limit their accessibility. This creates a need for lighter models that retain much of the LLM's intelligence but can be deployed efficiently. Knowledge distillation is one such technique that enables training smaller student models to replicate the behavior of a larger teacher model.

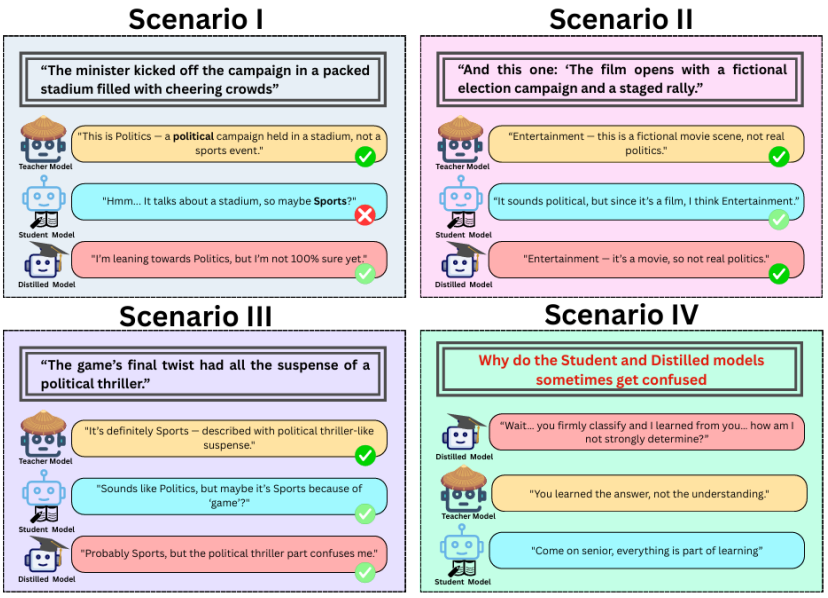

Figure 6:Explanatory diagram of text classification in teacher, student and distilled model

Knowledge distillation addresses this issue in a practical way by moving learned knowledge from a high-performing, computationally intensive teacher model to a smaller, more efficient student model. This teacher-student learning paradigm allows for the preservation of a considerable percentage of the teacher's knowledge while also enabling faster, lighter models suitable for real-world deployment. The ability to capture both hard label accuracy and subtle soft label distributions enables distillation models to generalize more well, particularly in multiclass contexts with overlapping semantics between classes.

In this study, we explore this concept for a zero-shot multiclass text classification problem using categories from real-world domains: entertainment, politics, and sports. We investigate the performance of teacher, student, and distilled models in terms of both accuracy and computational efficiency. Even though classification is challenging in the task agnostic distillation (see Figure 1) Additionally, traditional machine learning models such as Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks, and Naive Bayes are evaluated to provide baseline comparisons. These models were chosen after analyzing recent literature, which consistently ranks LSTM, CNN, SVM, and Naive Bayes as among the best for text categorization tasks. The comparison to these baselines allows us to evaluate not only the raw performance gains, but also the practical benefits of applying knowledge distillation to real-world classification problems.

LITERATURE REVIEW

This study builds on several foundational concepts in natural language processing and machine learning. Knowledge distillation (Pangakis & Wolken, 2024) is the process of transferring the generalization ability of a large model (teacher) into a smaller, more efficient model (student) by mimicking the teacher's output distributions. Historically, prior to the era of LLMs the knowledge distillation techniques primarily concentrated on transferring knowledge from complex, often complicated neural networks to more compact and efficient networks. After the arrival of transformer-based architectures, notably BERT (Bidirectional Encoder Representations from Transformers), GPT, and its variations, knowledge distillation approaches became popular in the field of Natural Language Processing (NLP). The current phase of knowledge

distillation in LLMs switches the focus beyond simple architectural model compression to knowledge elicitation and richer transfer of knowledge.

Knowledge Distillation (KD) has emerged as a powerful technique for model compression, where a smaller student model learns to mimic a larger teacher model. Extensive surveys explore KD in diverse domains including computer vision, graphs, large language models (LLMs), and federated learning. The teacher–student paradigm remains central to KD's effectiveness (Hu et al., 2023). KD for LLMs has attracted increasing attention due to computational demands, with works exploring data-efficient methods (Li et al., 2024), synthetic data, domain-specific distillation, and privacy-preserving data augmentation. Multi-aspect distillation frameworks and performance-guided strategies improve task alignment, particularly in text classification. Logic-based distillation techniques show promise in planning and decision-making (Chen et al., 2024). Approaches like MiniLLM provide scalable methods for LLM distillation (Gu et al., 2023), while sparse logit sampling(Zaidi et al., 2025) and smoothed distillation target hallucination and efficiency. Mitigating KL divergence limitations and enabling knowledge transfer across modalities such as provenance analysis (Zuo et al., 2025) and graphs broaden KD applications. KD also supports specialized tasks including anomaly detection and automatic scoring. Compression techniques like quantization-aware training (Liu et al., 2023), sub-4-bit distillation (Du et al., 2024), and pruning are essential for efficient deployment. Studies also assess compression's impact on parametric knowledge. Multi-teacher strategies have been explored for tasks such as reasoning-based machine reading comprehension. Furthermore, federated KD frameworks are evolving to address privacy and communication constraints. With increasing demand for scalable, generalizable, and low-resource NLP solutions, KD continues to be a critical research area across domains and architectures.

Table 1: Comparison of related works with our research

Previous Study	Primary Objective	Difference from Our Research
DistilBERT(Sanh et al., 2019)	Reduce BERT model size while retaining most of its performance on general NLP tasks.	Focuses on task-specific zero-shot text classification rather than general-purpose NLP.
TinyBERT(Jiao et al., 2019)	Knowledge distillation for efficient transformer models, both task-specific and general-purpose.	Evaluates zero-shot classification ability of distilled models without extensive task-specific fine-tuning.
PGKD(Di Palo et al., 2024)	Active-learning KD where task-specific student models are trained using LLM-generated hard negatives.	Employs few-shot, task-specific distillation with both hard and soft labels, without active learning, to balance accuracy and efficiency in resource-limited settings.
KDH-MLTC(Song et al., 2024)	Applies KD for multi-label medical text classification using PSO and sequential fine-tuning to boost student performance.	Directly transfers teacher knowledge to the student via few-shot, task-specific distillation without further fine-tuning.

Text classification is a fundamental operation in natural language processing that assigns specific groups to new text based on its semantic content. Multiclass classification (Yuan et al.) introduces further challenges by requiring models to learn nuanced differences between categories, particularly when categories may have overlapping language styles or semantics. This complexity grows when categories have comparable terminology or when textual expressions are ambiguous, as in sectors such as entertainment, sports, and politics. Our approach emphasizes the synergy between soft and hard labels, as well as the use of both semantic and task-specific evaluation metrics.

Knowledge Elicitation (Du et al., 2025) is the process of extracting useful and feasible knowledge from expert sources. Knowledge elicitation ensures that the knowledge transferred is not simply shallow imitation but rather reflects the teacher's profound language and semantic understanding.

Hard labels are true category labels assigned to each data sample, which are commonly expressed as discrete class identifiers. In contrast, soft labels, which include probability distributions or logits, offer richer

learning signals than hard labels alone, making them especially valuable for distillation. By combining both hard and soft labels during training, models can learn to match ground-truth assignments while also mimicking the teacher's more sophisticated prediction distributions. Logits are the raw, unnormalized output values produced by a neural network prior to the use of any activation function, such as SoftMax. They reflect the model's internal confidence scores for each class. During distillation, logits from the teacher model are utilized to generate soft labels that assist the student model in learning not just the proper class but also the relative relationships across classes.

Divergence-based loss functions, such Kullback-Leibler (KL) divergence (Wu et al., 2024), compute the difference between the teacher's soft label distribution and the student's projected distribution. By reducing divergence during training, the student model learns to closely replicate the teacher's behavior, integrating its understanding into a smaller, more efficient model. Table 1 summarizes recent studies on knowledge distillation techniques for task-specific text classification.

METHODOLOGY

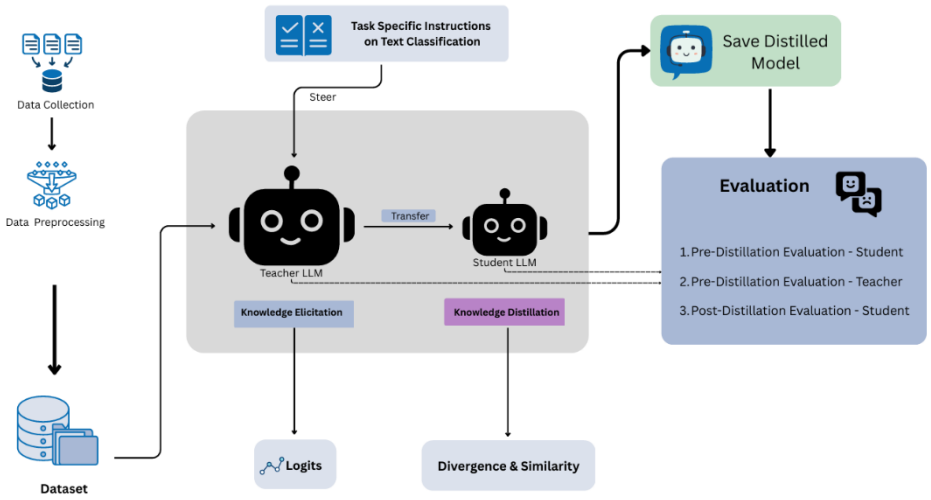

Figure 7: Methodology

Our research is to address the gap of making a task specific distilled model which should perform better than the student model and avoid hallucinations. We curated a dataset comprising text samples labeled under three categories: entertainment, sports, and politics. After initial preprocessing such as text normalization and cleaning, the dataset was prepared for model evaluation and training phases. As the methodology diagram shows, the whole methodology is divided into several stages, such as evaluation of teacher and student models, knowledge elicitation, knowledge distillation, and traditional baseline comparisons.

In the first phase, the preprocessed dataset was fed into the student model for preliminary evaluation. The student model used in this research is the LLaMA 3.1:8B model. During this step, the student model was given each input text sample without its corresponding label. The model made predictions using a prompt-based instruction style, with each sentence assigned to one of three target categories: sports, politics, or entertainment. The forecasts were then compared to the original ground truth labels. Due to computational restrictions, only 50 samples from CNN/DailyMail dataset were used for this Evaluation. Prior to distillation, performance indicators including as accuracy, precision, recall, F1-score, and confusion matrix were used to evaluate the student model's classification capability.

The same evaluation technique was used on the teacher model, which is the larger LLaMA 3.1: 70B model. The teacher model used the same prompt-based classification function as the student model. Predictions were created, and classification metrics were calculated by comparing the predicted labels to the actual

labels. Because of its larger parameter size, the teacher model performed better with complicated and semantically confusing examples.

After acquiring baseline results from both student and teacher models, the distillation step proceeded. During this stage, the labeled dataset consists of 3,000 samples from the CNN/DailyMail dataset was run through the teacher model to extract soft label outputs in the form of logits, which represented the teacher's confidence distribution across the three categories for each sample. The logits were then employed as supervisory signals during distillation training.

A divergence-based loss function was used to retrain the student model with both hard labels (ground truth) and soft labels (teacher logits). The distillation loss was estimated using a weighted combination of cross-entropy loss (on hard labels) and Kullback-Leibler divergence loss (on soft logits), allowing the student model to better approach the teacher's decision boundaries.

$$L_{KD} = \alpha L_{CE}(y_{pred}, y_{true}) + (1 - \alpha)T^2 \, KL(P_T \,||\, P_S)$$

Here, L_{CE} represents the cross-entropy loss between the student's predicted tokens and the ground-truth summaries, while KL(•) defines the divergence between the softened output distributions of the teacher (P_T) and student (P_S), with T as the temperature parameter that controls the softness of the probability distributions. The hyperparameter α balances the contributions of these two loss components. The combined loss function allows the student model to learn from both the real labels and the nuanced information included in the teacher's output probabilities, resulting in better semantic comprehension as well as improved predictions.

The distillation training approach allowed the student to acquire more information from the teacher while remaining lightweight and efficient. Following training, the distilled model was preserved for further review. The saved distilled student model was then evaluated with the same test data and standards as the initial student and teacher models. The distilled model's performance was evaluated by comparing its predictions to the ground-truth labels and again computing accuracy, precision, recall, and F1-score measures. In addition, average processing times were measured to compare the response times of these models. This allowed for direct comparisons between all models (see Figure 3).

In addition to the teacher-student distillation architecture, classic machine learning methods were used to set performance benchmarks. The models used are: LSTM, SVM, CNN and Naive Bayes. The dataset was divided into an 80:20 train-test ratio to ensure a balanced evaluation of the models. Prior to training, preprocessing processes such as tokenization, vectorization, and padding were used to turn the raw text input into formats suited for each method.

The Support Vector Machines (SVM) model employed TF-IDF to turn the text into feature vectors for classification. The Long Short-Term Memory (LSTM) model was trained on tokenized and padded word sequences, with embeddings representing the text inputs. The Naive Bayes (MultinomialNB) classifier also trained on TF-IDF features. To capture text patterns, the Convolutional Neural Network (CNN) model used tokenized and padded sequences that were fed through convolutional layers. The traditional models were evaluated using the same set of classification criteria as the distilled and LLM-based models, which included accuracy, precision, recall, and F1-score. This extensive examination provides useful comparative insights into the success of knowledge distillation vs traditional classification methodologies, emphasizing the distilled model's balance of computational efficiency and predictive performance.

RESULTS AND DISCUSSION

The teacher model achieved an accuracy of 55.3%, precision of 51.1%, recall of 58.0%, and F1-score of 56.7%, setting an upper benchmark in classification performance, but took a large amount of processing time. Its exceptional performance can be due to its large parameter set, which enables it to capture complicated linguistic patterns and contextual connections. However, this advantage comes at the expense of high computing requirements, rendering it unsuitable for implementation in real-time or resource-constrained contexts. The average reaction time for the teacher model was roughly 6-7 seconds per sample, whereas the student and distilled models were around 2-3 seconds per sample.

The student model classified quickly but performed poorly, particularly in semantically confusing scenarios. It struggled with classifications that included small overlaps between categories, such as texts

citing sports events in political contexts or technology concerns embedded inside entertainment tales. This is demonstrated in its relatively poor accuracy of 31.0%, precision of 29.0%, recall of 21.0%, and F1-score of 35.5%. The student model's lightweight design hindered its capacity to handle complex circumstances requiring fine-grained contextual comprehension.

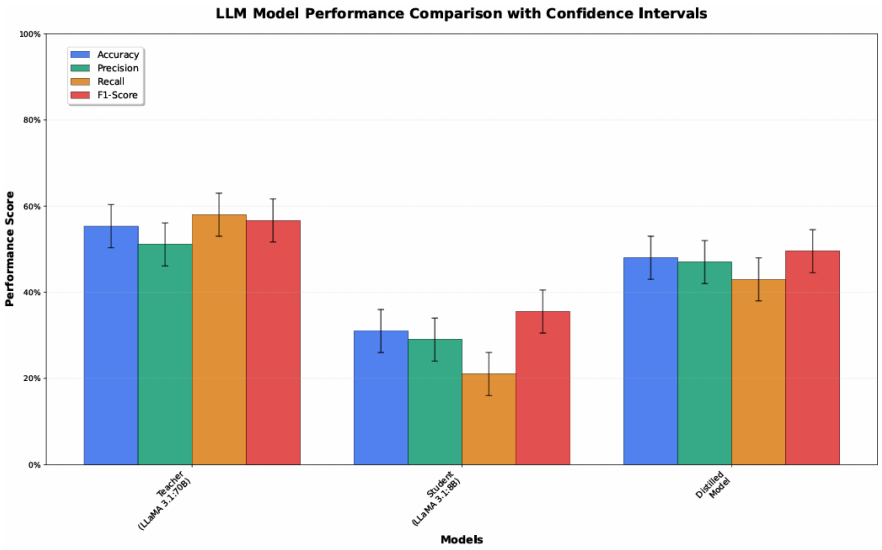

Figure 8: Model Performance Comparison of LLM models in Text Classification:

The distilled model demonstrated a solid balance, providing performance comparable to the teacher while being significantly faster. It achieved an accuracy of 48.0%, precision of 47.0%, recall of 43.0%, and F1-score of 49.5%. By adding the teacher's soft label distributions generated during distillation, the student model learnt to understand nuanced correlations between courses even when explicit cues were restricted. This resulted in better generalization, especially in borderline circumstances when hard labels alone may not offer adequate learning signals. The distilled model regularly displayed improved adaptability across domains, effectively retaining the teacher's reasoning abilities while minimizing computational footprint (see Figure 3).

Table 2:Performance Comparison of Teacher, Student, Distilled, and Traditional Models

Model	Accuracy	Precision	Recall	F1 Score
Teacher (LLaMA 3.1:70B)	0.5534	0.5111	0.5801	0.5668
Student (LLaMA 3.1:8B)	0.3100	0.2900	0.2101	0.3552
Distilled Model	0.4800	0.4700	0.4300	0.4952
LSTM	0.6770	0.6900	0.6055	0.6385
CNN	0.6100	0.7286	0.6300	0.6800
SVM	0.8010	0.8250	0.7200	0.7971
Naive Bayes	0.8440	0.8850	0.8600	0.8971

Traditional models like LSTM, CNN, and SVM fared well across multiple categories. Naive Bayes and SVM performed well on clearly separable samples with Naive Bayes reaching an accuracy of 84.4%, precision of 88.5%, recall of 86.0%, and F1-score of 89.7%, and SVM closely following with an accuracy of 80.1%, precision of 82.5%, recall of 72.0%, and F1-score of 79.7%. LSTM and CNN generated reasonable results by detecting sequential and local patterns in the text. However, these models demonstrated shortcomings when dealing with semantically ambiguous scenarios involving overlapping features between categories. In comparison, the distilled model struck a good mix between accuracy and computational efficiency while efficiently absorbing knowledge from the teacher model. While the

distilled model did not outperform all traditional models, it did show greater generalization over the initial student model, particularly in complicated classification cases. The detailed comparative results of all models are presented in Table 2.

Our findings indicate that task-specific distillation effectively compresses the teacher's knowledge into a lightweight model while maintaining relevant classification performance. The distilled model outperforms the student model and competes with standard models for domain-specific nuances, as evidenced by qualitative error analysis. Furthermore, analysis of misclassified instances revealed that the distilled model was more likely to reflect the teacher's conclusion when several semantic signals were present, suggesting its ability to grasp fine-grained distinctions across overlapping categories.

CONCLUSION

We developed a knowledge distillation framework tailored for multiclass text classification in entertainment, sports, and politics. Our **distilled model** combines the best of both worlds: the teacher's reasoning ability and the student's efficiency. This framework enables effective implementation in real-world applications with resource constraints. Beyond its immediate application, this approach provides a scalable solution that may be applied to a variety of topics and languages. The distilled model effectively captures nuanced semantic patterns by using both hard and soft labels, making it suited for challenging classification applications. Our methodology shows that knowledge distillation not only decreases computing needs but also maintains meaningful task-specific performance, indicating that it has the potential for widespread use in a variety of NLP applications. The performance of language models in any NLP task is comparatively low consider to traditional models which learns from the hard label. But Language models itself it is classifying in context of knowledge it has. So, task specific wise the model performance is low. But as task agnostic it is better in concurrent scenario. Further it can extend to expand the classification knowledge to the large language models as traditional models.

LIMITATIONS

Our work is currently limited to three domains and single-language datasets. The distilled model, while effective, still depends on the quality of the teacher model's outputs. Research is conducted only for the LLaMa model yet, but we had 4 different models in our pipeline. Due to resource constrained we developed one only.

Additionally, the semantic alignment was limited to textual similarity metrics; incorporating structural or contextual features could enhance performance further. Future work could explore multilingual classification, multimodal inputs, and further improvements in few-shot learning. Since our study largely focused on zero-shot classification, we intend to further our analysis into one-shot and few-shot learning contexts to better understand the benefits of knowledge distillation with limited labeled data sets. While the distilled model outperforms the student model, it falls short of the teacher model. This discrepancy may indicate insufficient fine-tuning or limited task adaptation. Due to resource constraints, we could not fully fine-tune the large language models, possibly limiting their effectiveness.

REFERENCES

1. Chen, D., Zhang, S., Gao, F., Zhuang, Y., Tang, S., Liu, Q., & Xu, M. (2024). Logic Distillation: Learning from Code Function by Function for Planning and Decision-making. *arXiv preprint arXiv:2407. 19405*.
2. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & Others. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, 2019). *Bert: Pre-training of deep bidirectional transformers for language understanding* Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),
4. Di Palo, F., Singhi, P., & Fadlallah, B. (2024). Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale. *arXiv preprint arXiv:2411. 05045*.

5. Du, D., Zhang, Y., Cao, S., Guo, J., Cao, T., Chu, X., & Xu, N. (2024). Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. *arXiv preprint arXiv:2402. 10631*.
6. Du, Y., Sun, Z., Wang, Z., Chua, H., Zhang, J., & Ong, Y.-S. (2025, 2025). *Active large language model-based knowledge distillation for session-based recommendation* Proceedings of the AAAI Conference on Artificial Intelligence,
7. Gu, Y., Dong, L., Wei, F., & Huang, M. (2023). MiniLLM: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.
8. Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., & Liu, X. (2023). Teacher-student architecture for knowledge distillation: A survey. *arXiv preprint arXiv:2308. 04268*.
9. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.
10. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.
11. Li, J., Nag, S., Liu, H., Tang, X., Sarwar, S., Cui, L., Gu, H., Wang, S., He, Q., & Tang, J. (2024). Learning with Less: Knowledge Distillation from Large Language Models via Unlabeled Data. *arXiv preprint arXiv:2411. 08028*.
12. Liu, Z., Oguz, B., Zhao, C., Chang, E., Stock, P., Mehdad, Y., Shi, Y., Krishnamoorthi, R., & Chandra, V. (2023). Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305. 17888*.
13. Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, *14*(1), 9603.
14. Pangakis, N., & Wolken, S. (2024). Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels. *arXiv preprint arXiv:2406. 17633*.
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & Others. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
16. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
17. Song, Y., Zhang, J., Tian, Z., Yang, Y., Huang, M., & Li, D. (2024). LLM-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402. 16515*.
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
19. Vörös, T., Bergeron, S. P., & Berlin, K. (2023, 2023). *Web content filtering through knowledge distillation of large language models* 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),
20. Wu, T., Tao, C., Wang, J., Yang, R., Zhao, Z., & Wong, N. (2024). Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404. 02657*.
21. Yuan, B., Chen, Y., & Zhang, Y. Label-Guided Self-Knowledge Distillation for Multi-Class Text Classification. *Available at SSRN 5081643*.
22. Zaidi, M. A., Kedia, A., Ahn, J., Kwon, T., Lee, K., Lee, H., Lee, J., & Others. (2025). Sparse Logit Sampling: Accelerating Knowledge Distillation in LLMs. *arXiv preprint arXiv:2503. 16870*.
23. Zuo, F., Rhee, J., & Choe, Y. R. (2025). Knowledge Transfer from LLMs to Provenance Analysis: A Semantic-Augmented Method for APT Detection. *arXiv preprint arXiv:2503. 18316*.