Multi-Teacher Knowledge Distillation with Clustering-Based Sentence Pruning for Efficient Student Models

Anonymous ACL submission

Abstract

Transformer-based encoder models such as BERT and RoBERTa perform well on NLP tasks but are computationally intensive for deployment. We propose **Clustering-Based Knowledge Distillation with Sentence Pruning**, a novel framework that combines multi-teacher distillation with structure-aware sentence selection to improve student model efficiency. Our method integrates teacher outputs via validation-aware ensembling and prunes redundant sentences using semantic similarity and TF-IDF-based scoring. Experiments across GLUE, AG News, and PubMed RCT demonstrate that our method consistently enhances student model performance, achieving 95.4% accuracy on SST-2, the highest accuracy on AG News (91.14%) and PubMed RCT (78.00%), and improved accuracy on RTE through sentence pruning. Ablation studies confirm the effectiveness of jointly applying clustering and pruning. Our framework offers a practical and scalable solution for deploying compact models in resource-limited environments.

1 Introduction

Transformer-based pre-trained models, such as **BERT**, **RoBERTa**, and **GPT**, have set new standards in NLP tasks and achieved state-of-the-art performance across classification, inference, and generation (Koroteev, 2021; Delobelle et al., 2020; Achiam et al., 2023). However, their substantial computational requirements pose challenges for real-world deployment, particularly in low-power and constrained computing environments(Jiao et al., 2020). To address this challenge, **Knowledge Distillation (KD)** has been widely adopted as an effective model compression technique that transfers knowledge from a large **teacher model** to a smaller **student model**, enabling efficient inference while maintaining high performance. Despite its effectiveness, conventional knowledge distillation (KD) methods face several limitations. While a variety of KD techniques—including those that align intermediate representations, such as MiniLM (Wang et al., 2020), TinyBERT, and CoDIR (Zhang et al., 2023)—have been proposed to enrich the transfer process beyond output distributions, many of these approaches still struggle to effectively capture **inter-sentence dependencies**. These aspects are particularly crucial for complex NLP tasks such as natural language inference and summarization (Wei et al., 2024).

Moreover, most existing KD frameworks adopt a **single-teacher** paradigm, which inherently limits the diversity and richness of knowledge imparted to the student (Pham et al., 2023). This lack of heterogeneity in supervision can lead to reduced generalization, especially when the teacher model fails to cover all linguistic variations relevant to the task. Furthermore, transferring knowledge directly from a large, complex teacher model can introduce **noisy or overly sophisticated signals**, which may overwhelm the capacity of compact student models and hinder effective learning (Yuan et al., 2024).

Distillation Method	Teacher Acc. (%)	Student Acc. (%)	Discrepancy Acc. (%)
AVER-Student	81.41	64.75	–16.66
EBKD-Student	**81.57**	64.66	–16.91
MMKD-Student	79.13	**64.87**	–14.26

Table 1: Comparison of teacher and student accuracies across distillation methods: AVER (Fukuda et al., 2017), EBKD (Kwon et al., 2020), and MMKD (Wei et al., 2024).

Despite the use of ensemble teachers, we observe a noticeable discrepancy between teacher ensemble accuracy and the final performance of the student model. As summarized in Table 1, although the EBKD strategy achieves the highest ensemble accuracy (81.57%), it yields a lower student accuracy (64.66%) compared to the MMKD method (64.87%). Interestingly, MMKD—despite being associated with the lowest ensemble accuracy

(79.13%)—outperforms other methods in terms of student generalization.

This result indicates that a higher ensemble teacher accuracy does not necessarily translate to improved student performance. In particular, the MMKD approach, which individually distills knowledge from multiple teacher models rather than aggregating their predictions, appears to better preserve the diversity of knowledge. Such diversity facilitates more robust learning signals, thereby enhancing the generalization ability of the student model. These findings highlight that the *methodology of ensemble integration* significantly affects the quality of distilled knowledge, suggesting that selecting appropriate ensemble-distillation schemes is critical for maximizing student performance.

To overcome these limitations, we propose **Clustering-based Knowledge Distillation with Sentence Pruning Processing**, a novel framework that enhances knowledge transfer by integrating multiple teacher models while refining the input representation through sentence-level pruning. Our method utilizes Clustering-based modeling of inter-sentence relationships, which aggregates knowledge from multiple teacher models to enhance robustness and diversity while modeling inter-sentence relationships through clustering-based representation. This approach effectively retains essential information, optimizing the student model's learning process. This work makes the following methodological contributions:

- We present a **clustering-based pruning method** that selects key sentences using TF-IDF and cluster centrality within semantic groups.
- We design a unified framework that integrates **multi-teacher distillation** with pruning to enhance efficiency and robustness.
- We enable **efficient lightweight student training** by combining a performance-weighted teacher ensemble and selective input pruning.

2 Related Work

2.1 Knowledge Distillation

Knowledge Distillation (KD) is a widely adopted model compression technique that facilitates knowledge transfer from a large, high-capacity teacher model to a smaller, lightweight student model (Gu et al., 2024). The core idea is to guide the student using soft targets—typically the output probability distributions or intermediate representations—produced by the teacher.

These soft labels encode semantic similarity among classes, offering richer signals than hard labels (Gao, 2023). To smooth the transfer process, temperature scaling is often used to soften logits, helping the student mimic the teacher's confidence distribution more effectively. Beyond output alignment, KD has expanded to include intermediate-layer feature matching, where the student aligns its hidden states with those of the teacher (Haidar et al., 2021; Zhang et al., 2024). This enables the student to benefit from hierarchical abstraction learned by the teacher. Recent research has introduced extensions such as attention-guided layer alignment (Passban et al., 2021), structured hidden-state distillation (Zhou et al., 2022), and relational knowledge selection (Xu et al., 2020), further enhancing transfer effectiveness.

KD has proven successful across diverse NLP tasks—including classification, question answering, and inference—by enabling smaller models to inherit generalization capabilities from larger ones (Song et al., 2022; Yuan et al., 2021). Reinforcement learning-based KD frameworks (Qiu et al., 2022; Hong et al., 2021) and adaptive supervision strategies (Du et al., 2020) have also emerged, offering dynamic and data-aware distillation paradigms. These developments position KD as a flexible and powerful framework for training compact yet capable models, laying the foundation for broader ensemble distillation techniques discussed next.

2.2 Limitations of Existing Approaches

While conventional knowledge distillation (KD) has proven effective in compressing large models, it suffers from several notable limitations. First, **single-teacher distillation** restricts the diversity of knowledge transferred to the student model, often resulting in limited generalization, particularly in linguistically diverse tasks (Yuan et al., 2021; Wu et al., 2022). To overcome this, **ensemble-based KD** has been introduced, wherein multiple teacher models provide more comprehensive and diverse supervision. However, naively aggregating outputs—such as averaging logits—can lead to **conflicting or redundant knowledge**, which may confuse or overwhelm the student (Shao and Chen, 2023). Moreover, such aggregation fails to account for the varying reliability of individual teachers, especially across different input distributions. Recent studies have proposed adaptive weighting and reinforcement learning-based teacher selection mechanisms (Du et al., 2020; Qiu et al., 2022), yet these still struggle to **filter out noisy or overly complex signals**. This is particularly problematic when the student model has limited capacity, as it cannot effectively absorb dense or conflicting supervision (Fan et al., 2021; Yuan et al., 2021). As a result, existing ensemble KD methods remain suboptimal in balancing supervision diversity with the student's representational limits. These challenges underline the need for a **more structured and selective approach** to ensemble knowledge distillation—one that not only aggregates diverse knowledge but also prunes irrelevant or noisy content prior to student training.

3 Method

As illustrated in Figure 1, our proposed framework comprises three core components. First, we employ a **validation-aware ensemble distillation** strategy (LR-Dev-Ensemble), where multiple teacher models are combined using logistic regression trained on the validation set, allowing the framework to weigh each teacher's output based on its generalization ability. Second, a **clustering-based sentence pruning module** analyzes the sentence similarity structure, clusters semantically related sentences based on cosine similarity, and dynamically prunes redundant or low-importance sentences using TF-IDF-based thresholds within each cluster. Finally, the **student model is trained** using both soft targets from the ensembled teachers and hard labels from the ground truth, optimized via a combined loss function that integrates KL divergence and cross-entropy. This integrated design ensures that the student receives both diverse and compressed knowledge, improving generalization while reducing computational cost.

3.1 Ensemble-Based Knowledge Distillation

We adopt a single ensemble strategy, referred to as **LR-Dev-Ensemble**, to combine the outputs of multiple teacher models. **LR-Dev-Ensemble** is a validation-aware ensemble strategy that trains a logistic regression model on development data to learn optimal weights for combining teacher outputs. Unlike uniform or fixed-weight averaging, it dynamically reflects each teacher's reliability, offering a more discriminative and generalizable soft target for student training. In this approach, a logistic regression model is trained using the validation set outputs of each teacher model to learn the optimal combination weights. These weights reflect the generalization ability of each teacher and are used to form a weighted ensemble distribution. This weighting mechanism enables more effective knowledge transfer, as higher weights are assigned to teachers with better validation performance.

Formally, for a given input x, let the softmax output of the i-th teacher model be $P_{teacher_i}(y|x)$. The final ensemble distribution is computed as:

$$P_{ensemble}(y|x) = \sum_{i=1}^{N} \alpha_i \cdot P_{teacher_i}(y|x), \quad (1)$$

where α_i denotes the learned weight for teacher i, subject to $\sum_{i=1}^{N} \alpha_i = 1$.

The ensemble output $P_{ensemble}(y|x)$ is then used as a soft target to train the student model by minimizing the Kullback–Leibler (KL) divergence between the student output and the ensemble distribution.

This validation-aware weighted ensemble approach enables more robust and efficient knowledge transfer, as it down-weights less reliable teachers and avoids misleading or noisy supervision signals. Consequently, the student model benefits from a more informative and generalizable training signal.

3.2 Clustering-based Sentence Pruning

Ensemble distillation provides a comprehensive and nuanced representation of knowledge; however, directly utilizing outputs from multiple teacher models often introduces redundancy. This increases computational overhead and may degrade the training efficiency of the student model.

To address this, we propose a **clustering-based sentence pruning** strategy that systematically removes redundant or less informative sentences while preserving semantic relevance.

As shown in Figure 1, the pruning process begins by modeling pairwise sentence similarities, where each sentence is compared based on cosine similarity between their embeddings:

$$w_{ij} = \cos(\mathbf{E}_{v_i}, \mathbf{E}_{v_j}) = \frac{\mathbf{E}_{v_i} \cdot \mathbf{E}_{v_j}}{\|\mathbf{E}_{v_i}\| \|\mathbf{E}_{v_j}\|} \quad (2)$$

Next, we apply a clustering algorithm to group semantically similar sentences. The purpose of clustering is not only to group related sentences

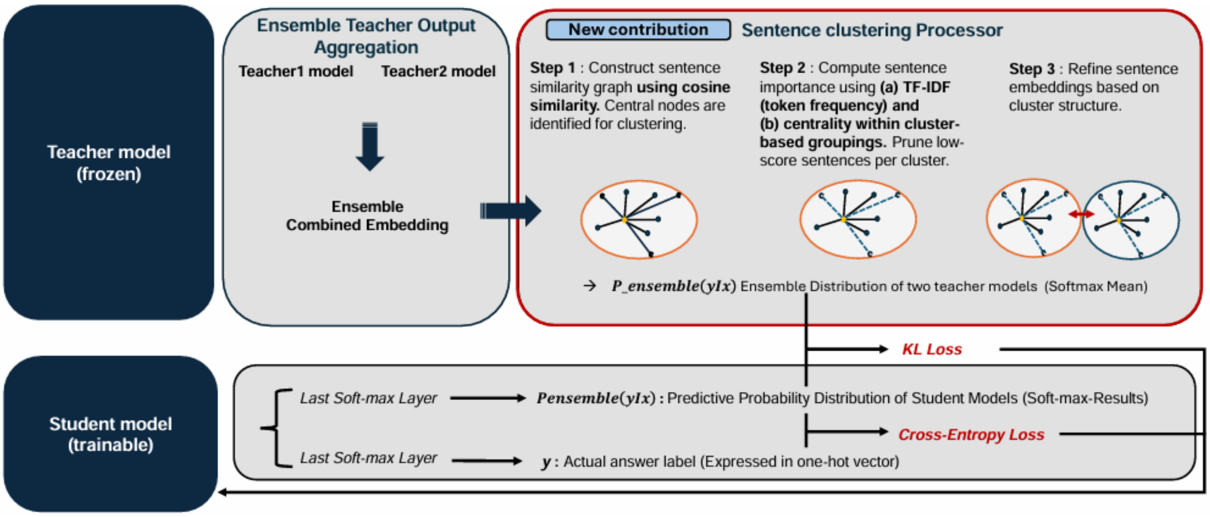

Figure 1: Overview of the Multi-Teacher Knowledge Distillation Framework with Clustering-Based Sentence Pruning

but also to constrain importance scoring within semantically coherent subsets. Rather than treating clustering as a standalone step, we leverage it to define *context-aware local neighborhoods*, enabling more precise computation of sentence importance relative to local context. This localized perspective helps our method avoid global importance bias and improves structural preservation during pruning.

We compute the importance of each sentence through a composite scoring mechanism that reflects both lexical frequency and structural centrality. Formally, the importance score $I(v_i)$ of sentence v_i is defined as:

$$I(v_i) = \lambda \cdot TFIDF(v_i) + (1 - \lambda) \cdot Centrality(v_i) \quad (3)$$

Here, $TFIDF(v_i)$ denotes the aggregated TF-IDF score of words in sentence v_i, and $Centrality(v_i)$ is measured as the cosine similarity between the sentence embedding and the centroid of its cluster:

$$Centrality(v_i) = \cos(\mathbf{E}_{v_i}, \mathbf{c}_k), \mathbf{c}_k = \frac{1}{|C_k|} \sum_{v_j \in C_k} \mathbf{E}_{v_j} \quad (4)$$

This formulation ensures that only semantically meaningful and structurally important sentences are retained within each cluster, thereby improving the efficiency of downstream student training while preserving essential contextual information.

After pruning, we retain the pre-computed sentence embeddings of the selected sentences—originally generated from the teacher encoder—and feed them into the student model as inputs. This preserves structural and semantic consistency between the teacher's supervision signals and the student's internal representation. Consequently, the student learns from a compact, structure-aware representation distilled from diverse teacher outputs.

3.3 Student Model Training

In our framework, the student model is trained using both soft labels generated by the LR-Dev-Ensemble strategy and hard labels from the ground truth. As introduced in Section 3.1, LR-Dev-Ensemble learns optimal weights over multiple teacher models based on validation performance, yielding a soft target distribution that reflects the relative strengths of each teacher. This enhances supervision quality by providing a more robust and generalizable signal for student training. The student model is optimized using a composite loss function that combines Cross-Entropy (CE) loss and Kullback-Leibler (KL) divergence, weighted by a coefficient $\lambda \in [0, 1]$:

$$\mathcal{L}_{total} = (1 - \lambda) \cdot \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{KL} \quad (5)$$

Before loss computation, the student input is refined via a clustering-based sentence pruning module, which filters redundant or noisy sentences to reduce input length while preserving semantic relevance. Note that pruning is applied only on the student-side inputs, while teacher soft targets are computed from the original, unpruned sequences. This decoupled design allows the student to benefit from full teacher supervision with minimal input overhead. Pruning operates *within each semantic cluster*, preserving local discourse structure. Sentence embeddings are first used to compute pairwise cosine similarities, from which we calculate a threshold $\tau = \mu + \alpha \cdot \sigma$, where μ and σ are the mean and standard deviation of all similarity scores. Sentence pairs with similarity above this threshold are grouped together, and clusters are formed by identifying sets of mutually similar sentences.

As shown in Table 8, our model maintains robust performance across a range of cluster configurations. This is attributed to our scoring mechanism, which balances lexical importance (TF-IDF) and structural centrality. Together, LR-Dev-Ensemble supervision and structure-aware pruning enable efficient and effective training of compact student models, improving both inference speed and generalization.

4 Experiments

4.1 Experimental Setup and Data Statistics

We evaluate our method on six tasks from the GLUE benchmark (Wang et al., 2018), including RTE (textual entailment), QQP (paraphrase detection), QNLI (Rajpurkar et al., 2016) (QA-based inference), SST-2 (Socher et al., 2013) (sentiment analysis), MNLI-m (Williams et al., 2017) (multi-genre entailment), and MRPC (Dolan and Brockett, 2005) (paraphrase detection). We further test on AG News (Zhang et al., 2015), a four-class topic classification task, and PubMed RCT (Dernoncourt and Lee, 2017), a biomedical sentence classification dataset. Together, these tasks form a diverse benchmark for evaluating the generalization of our approach. Dataset statistics are shown in Table 2.

Dataset	#Train	#Dev	#Test
RTE	2,490	277	3,000
QQP	363,849	40,430	390,965
QNLI	104,743	5,463	5,463
SST-2	67,349	872	1,821
MNLI-m	392,702	9,815	9,796
MRPC	3,668	408	1,725
AG News	101,000	9,000	7,600
PubMed RCT	180,000	10,000	10,000

Table 2: Statistics of the datasets used in our experiments. In addition to standard GLUE tasks (e.g., RTE, QQP, QNLI), AG News and PubMed RCT are included for evaluating document classification and biomedical summarization respectively.

4.2 Baseline Models and Implementation Details

For evaluating our approach, we compared it against multiple baseline methods. Vanilla Knowledge Distillation (V-KD) (Hao et al., 2023) trains student models using a single teacher, such as BERT_{12} or RoBERTa_{12}. U-Ensemble Teacher(Yang et al., 2020), averages the outputs of all teacher models by assigning them equal weights. Rand-Single-Ensemble Teacher(Fukuda et al., 2017), randomly selects a teacher model for each mini-batch to generate soft targets for student training. W-Ensemble Teacher(Chebotar and Waters, 2016), applies pre-determined, fixed weights to each teacher model. LR-Ensemble Teacher employs a Logistic Regression-based approach to adaptively compute the optimal weights for teacher models. Depending on whether the weights are learned from the training set or the development set, the method is referred to as LR-Train-Ensemble and LR-Dev-Ensemble, respectively. For the teacher models, we fine-tuned widely-used transformer architectures, including BERT_{12} and RoBERTa_{12}, where the subscript 12 denotes that each model consists of 12 transformer layers. To construct student models, we utilized simplified versions of BERT, incorporating 4 and 6 transformer layers, denoted as BERT_4 and BERT_6, respectively. This aligns with the methodology presented in Patient KD (Sun et al., 2019).

4.3 Experimental Setup

Our experiments followed the Patient KD framework. The student models, BERT_4 and BERT_6, were initialized using the bottom 4 and 6 layers of BERT-Base. Their distillation process involved tuning hyperparameters such as temperature T values $\{5, 10, 20\}$, loss balance coefficients α $\{0.2, 0.5, 0.7\}$, and γ values $\{0.3, 0.5, 0.7, 0.9\}$, optimized based on the development set.

For fine-tuning the teacher models, we utilized publicly available pre-trained weights from BERT_{12} and RoBERTa. The training setup included learning rates of $\{1e-5, 2e-5, 5e-5\}$, a batch size of 32, a sequence length of 128, and 4 training epochs. The best-performing model was selected based on accuracy on the development set.

To enhance the distillation process, a logistic regression-based policy function was employed for teacher selection, optimized using Monte Carlo policy gradients (Williams, 1992).

4.4 Comparison to Baselines

Following pretraining, Knowledge Distillation (KD) and Teacher Selection (TS) models (Ye et al., 2020; Amara et al., 2022; Lee et al., 2023) were trained iteratively in an alternating manner.

Model (T=Teacher)	Params	FLOPs	Avg. GLUE Score
BERT-B (T)	109M	22.5B	80.6
BERT-L (T)	340M	110B	81.6
RoBERTa-B (T)	125M	40B	91.1
D6 ← BERT	67M	11.3B	77.5
D4 ← BERT	52M	7.6B	74.8
BERT-B ← B+L	109M	22.5B	82.5
D6 ← B+R	67M	11.3B	83.0
D4 ← B+L	52M	7.6B	72.0

Table 3: Summary of teacher/student models: model size, FLOPs, and average GLUE accuracy. **Abbreviations:** D6/D4 = DistilBERT with 6/4 layers, B+L = BERT-Base + BERT-Large, B+R = BERT-Base + RoBERTa-Base, (T) = Teacher.

Table 3 demonstrates that our proposed framework consistently delivers strong performance across a wide range of teacher-student configurations. The student model **D6** ← **B+R**, distilled from both BERT and RoBERTa, achieves the highest average GLUE score of 83.0, confirming the effectiveness of multi-teacher distillation. Even in cases with reduced model capacity, such as **D4** ← **B+L**, our method maintains competitive performance (72.0), showing that it generalizes well across various model sizes and teacher combinations. This highlights that existing ensemble-based distillation strategies offer meaningful performance improvements and serve as strong foundations for further enhancement.

4.5 Main Results

Table 4 compares multiple knowledge distillation strategies, highlighting differences in teacher selection and aggregation. Baseline methods such as Rand-Single-Ensemble and W-Ensemble adopt random or uniform teacher usage, while LR-Dev-Ensemble and Best-Single-Ensemble utilize dev-set-guided selection. MT-BERT-Ensemble employs joint training, and RL-KD variants leverage reinforcement learning with three reward types: prediction accuracy (reward1), logit similarity (reward2), and task-specific metrics (reward3).

Our method outperforms existing approaches on large-scale tasks such as MNLI-m (87.17) and SST-2 (95.4), demonstrating strong generalization. In particular, the method achieves the highest accuracy on AG News (91.14) and PubMed RCT (78.00), confirming the scalability of our sentence pruning and RL-KD strategy to long-text and document-level classification. Although performance on MRPC and RTE is slightly lower, this is primarily due to the limited size and semantic variability of these datasets, which constrain the effectiveness of reward-based teacher selection. Nonetheless, our method remains valid and robust, as it consistently improves performance on large-scale tasks and maintains competitive accuracy even under low-resource scenarios. The integration of sentence-level teacher representation further facilitates context-aware knowledge transfer.

Table 5 presents the impact of sentence pruning on accuracy and F1 score across three GLUE tasks: SST-2, RTE, and QNLI. The pruning process led to varying effects on model performance, with accuracy retention differing across tasks. In the SST-2 dataset, the pruning rate was 5.7%, resulting in a marginal decrease of 0.50% in accuracy and 0.34% in the F1 score, indicating that the model remained relatively robust to pruning. Conversely, in the RTE dataset, pruning led to a significant improvement in accuracy, increasing from 64.29% to 68.75% (+4.5%), with a corresponding F1 score increase of +2.6%. This suggests that pruning effectively removed non-informative sentences, thereby enhancing model performance. In contrast, for QNLI, which had a pruning rate of 31.7%, the accuracy decreased slightly by 0.62%, and the F1 score was reduced by 0.35%. These results indicate that while pruning improves computational efficiency, its impact on accuracy is task-dependent.

Table 6 compares the performance of various clustering methods on the MNLI-m dataset, including our proposed **Clustering-Based Sentence Pruning** method, as well as K-Means, Spectral, Agglomerative, Mean Shift, and Gaussian Mixture Model (GMM). Among all approaches, the Clustering-Based Sentence Pruning method achieves the highest classification accuracy (82%) and the best silhouette score (0.65), indicating superior overall performance in both task-specific and structural clustering metrics.

K-Means, a centroid-based algorithm that partitions data by minimizing within-cluster variance, shows relatively high accuracy (78%) but a lower silhouette score (0.58), suggesting weaker cohesion among clusters. Spectral Clustering, which leverages graph Laplacians and eigenvectors of similarity matrices, performs moderately due to its sensitivity to pairwise similarity noise.

Teacher	Student	Strategy	MNLI-m (Acc.)	MRPC (Acc.)	RTE (Acc.)	SST-2 (Acc.)	AG News (Acc.)	PubMed RCT (Acc.)
Rand-Single-Ensemble	BERT6	V-KD	80.7	77.7	61.7	90.6	87.2	72.1
W-Ensemble	BERT6	V-KD	77.2	81.1	62.1	90.6	86.3	73.4
LR-Dev-Ensemble	BERT6	V-KD	81.1	80.6	64.6	90.8	88.5	74.2
Best-Single-Ensemble	BERT6	V-KD	80.5	80.4	66.1	90.3	88.1	74.6
MT-BERT-Ensemble	BERT6	RL-KD	–	–	**75.7**	94.6	90.2	76.5
RL-KD (reward1)	BERT6	RL-KD	82.0	82.8	67.1	91.7	89.3	75.2
RL-KD (reward2)	BERT6	RL-KD	82.1	82.1	67.2	91.4	89.5	75.4
RL-KD (reward3)	BERT6	RL-KD	81.6	**83.3**	68.2	92.3	90.1	76.8
Our Method	BERT6	**RL-KD**	**87.17**	70.9	60.7	**95.4**	**91.14**	**78.00**

Table 4: Performance comparison with state-of-the-art knowledge distillation strategies using BERT6 as the student model across seven classification tasks. Our proposed Clustering-Based Knowledge Distillation with Sentence Pruning shows consistent improvement over strong KD baselines, particularly in document-level tasks (AG News, PubMed RCT).

Task	Prune Rate (%)	Acc Base	Acc Pruned	ΔAcc. (%)	F1 Base	F1 Pruned	ΔF1 (%)
SST-2	5.7	51.72	51.22	-0.50	39.27	38.93	-0.34
RTE	**32.8**	64.29	**68.75**	**+4.5**	53.46	**56.02**	**+2.6**
QNLI	31.7	44.32	43.70	-0.62	39.09	38.74	-0.35

Table 5: Impact of Sentence Pruning on Accuracy and F1 Score.

Clustering Method	Accuracy (%)	Silhouette Score
Clustering-Based Sentence Pruning (Ours)	**82**	**0.65**
K-Means Clustering	78	0.58
Spectral Clustering	75	0.52
Agglomerative Clustering	76	0.56
Mean Shift Clustering	71	0.51
Gaussian Mixture Model (GMM)	77	0.57

Table 6: Comparison of accuracy and silhouette score across different clustering methods on the MNLI-m dataset.

Agglomerative Clustering, a hierarchical bottom-up approach, produces stable but average results in both accuracy and silhouette score. Mean Shift, which shifts data points toward local density maxima, performs worse in both metrics, likely due to over-fragmentation in high-dimensional space. GMM, a probabilistic model that treats the data as a mixture of Gaussians, shows a balanced performance (77% accuracy and 0.57 silhouette score), but still falls short of our Clustering-Based Sentence Pruning Method.

Overall, the results highlight that our Clustering-Based Sentence Pruning method is more effective for sentence-level representation grouping in distillation tasks, providing both semantically coherent clusters and improved downstream accuracy.

Pruning Method	Accuracy (%)	Training Time (min)
No Pruning (Original)	84.52	**7.40**
Saliency-Based Pruning	81.67	7.05
Clustering-Based Sentence Pruning (Ours)	**83.91**	**7.35**
Entropy-Based Pruning	81.06	7.26

Table 7: Performance comparison of sentence pruning methods on the MNLI dataset. The proposed method combines TF-IDF scoring and cluster-based sentence centrality to prune redundant content.

Table 7 summarizes the evaluation results of various sentence pruning techniques applied to the MNLI dataset. The Original setting, which uses the full input text without pruning, achieves the highest accuracy of 84.52% and serves as the performance upper bound. However, it also incurs the longest training time (7.40 minutes), as it processes all sentences during model training.

In contrast, pruning-based methods reduce training time by selecting a subset of informative sentences. The proposed **Clustering-Based Sentence Pruning (Ours)** method achieves a competitive accuracy of 83.91%, while slightly increasing training time to 7.35 minutes compared to other pruning techniques. This marginal increase reflects the cost of more refined sentence selection via structural similarity and TF-IDF analysis, which enables the model to retain semantically meaningful content more precisely. **Saliency**- and **Entropy-based** methods show lower accuracies (81.67% and 81.06%, respectively), implying potential information loss due to reliance on local gradient signals or prediction uncertainty.

Number of Clusters	Matched Accuracy (%)	Mismatched Accuracy (%)
3	81.12	81.71
5	81.34	**81.91**
7	81.30	81.79
10	**81.36**	81.66

Table 8: Ablation study results on the MNLI dataset with varying cluster counts. The *matched* set consists of in-domain examples, while the *mismatched* set contains out-of-domain examples.

To examine the effect of cluster granularity in structure-aware knowledge distillation using sentence similarity, we conducted an ablation study by varying the number of clusters ({3, 5, 7, 10}). Table 8 presents evaluation results on the MNLI dataset, using both the matched set (in-domain) and the mismatched set (out-of-domain), which serve to assess generalization performance.

The student model demonstrated stable performance across settings, with matched accuracy ranging from 81.12% to 81.36%, and mismatched accuracy between 81.66% and 81.91%. While matched accuracy slightly improved with more clusters—peaking at 10 clusters—the best mismatched performance (81.91%) was observed at 5 clusters. This suggests that moderate clustering offers a trade-off between semantic granularity and generalizability. Fewer clusters may lead to under-separation of diverse sentences, while excessive clustering could reduce intra-cluster coherence.

These results highlight the importance of selecting an appropriate cluster count in structure-aware knowledge distillation using sentence similarity.

4.6 Ablation Study

Method	Accuracy (%)
Clustering + Pruning	**87.42**
Clustering Only	85.18
Pruning Only	83.26
No Processing	81.09

Table 9: Ablation study results on the MNLI dataset. Combining clustering and pruning yields the highest accuracy.

The results in Table 9 present the performance impact of different sentence processing strategies on the MNLI dataset. Notably, the *Clustering with Pruning* configuration achieves the highest accuracy of 87.42%, clearly outperforming all other baselines. This demonstrates that combining semantic-aware sentence selection (clustering) with redundancy reduction (pruning) leads to complementary effects that enhance model performance.

Comparatively, applying Clustering Only results in 85.18% accuracy, outperforming the Pruning Only (83.26%) setting. This suggests that semantic clustering contributes more to the model's generalization capability than structural pruning alone. Finally, the No Processing baseline achieves the lowest accuracy at 81.09%, highlighting the effectiveness of incorporating both clustering and pruning mechanisms into the knowledge distillation framework.

5 Conclusion

In this study, we proposed a **Clustering-Based Knowledge Distillation with Sentence Pruning** framework that combines **multi-teacher distillation** and **structure-aware pruning** to improve student model efficiency and generalization. Our method selectively filters redundant content using clustering and TF-IDF scoring, preserving key semantics. Experiments across tasks including **SST-2, RTE, QNLI, AG News, and PubMed RCT** show that our approach achieves strong accuracy with reduced inference cost. It also attains top performance on document-level tasks such as **AG News (91.14)** and **PubMed RCT (78.00)**. While minor drops occur on tasks like QNLI, the overall trade-off remains favorable. Our results highlight the framework's suitability for **resource-constrained deployment**, offering a scalable and effective strategy for compact model training.

6 Limitations

Although the proposed method demonstrates strong performance across diverse benchmarks, it exhibits comparatively lower accuracy on **MRPC** and **RTE** due to dataset-specific challenges. In MRPC, the task relies on fine-grained lexical overlap between sentence pairs, which can be inadvertently disrupted by pruning. RTE requires entailment decisions based on minimal context, often involving implicit reasoning, which may not be adequately captured through sentence-level clustering or teacher aggregation. These limitations indicate that **task-specific adaptations**, such as overlap-preserving pruning or external knowledge integration, may further improve performance on such datasets.

References

Josh Achiam et al. 2023. Gpt-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Ibtihel Amara et al. 2022. Ces-kd: curriculum-based expert selection for guided knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1901–1907. IEEE.

Yevgen Chebotar and Austin Waters. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: A dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases.

In *Third International Workshop on Paraphrasing (IWP2005)*.

Shangchen Du et al. 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In *Advances in Neural Information Processing Systems*, volume 33, pages 12345–12355.

Yang Fan et al. 2021. Learning to reweight with deep interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7385–7393.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Proceedings of Interspeech*, pages 3697–3701.

Minghong Gao. 2023. A survey on recent teacher-student learning studies. *arXiv preprint*, arXiv:2304.04615.

Yuxian Gu et al. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Md Akmal Haidar et al. 2021. Rail-kd: Random intermediate layer mapping for knowledge distillation. *arXiv preprint*, arXiv:2109.10164.

Zhiwei Hao et al. 2023. Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale. *arXiv preprint*, arXiv:2305.15781.

Zhang-Wei Hong, Prabhat Nagarajan, and Guilherme Maeda. 2021. Periodic intra-ensemble knowledge distillation for reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I*, pages 87–103. Springer International Publishing.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Mikhail V. Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. *arXiv preprint*, arXiv:2103.11943.

Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. 2020. Adaptive knowledge distillation based on entropy. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7409–7413. IEEE.

Hayeon Lee et al. 2023. A study on knowledge distillation from weak teacher for scaling up pre-trained language models. *arXiv preprint*, arXiv:2305.18239.

Peyman Passban, Qun Zhang, and Xuan Zhang. 2021. Alp-kd: Attention-aware layer projection for knowledge distillation. In *ACL*.

Cuong Pham, Tuan Hoang, and Thanh-Toan Do. 2023. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6435–6443.

Zengyu Qiu et al. 2022. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*.

Pranav Rajpurkar et al. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint*, arXiv:1606.05250.

Baitan Shao and Ying Chen. 2023. Decoupled knowledge with ensemble learning for online distillation. *arXiv preprint arXiv:2312.11218*.

Richard Socher et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Jie Song et al. 2022. Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, 31:3359–3370.

Siqi Sun et al. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint*, arXiv:1908.09355.

Alex Wang et al. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint*, arXiv:1804.07461.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *ACL*.

Jingxuan Wei, Yifan Gao, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2024. Sentence-level or token-level? a comprehensive study on knowledge distillation. *arXiv preprint*, arXiv:2404.14827.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Chuhan Wu et al. 2022. Unified and effective ensemble knowledge distillation. *arXiv preprint arXiv:2204.00548*.

Guodong Xu et al. 2020. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604, Cham. Springer International Publishing.

Ze Yang et al. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, pages 690–698.

Han-Jia Ye, Su Lu, and De-Chuan Zhan. 2020. Distilling cross-task knowledge via relationship matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12396–12405.

Fei Yuan et al. 2021. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14284–14291.

Mengyang Yuan, Bo Lang, and Fengnan Quan. 2024. Student-friendly knowledge distillation. *Knowledge-Based Systems*, 296:111915.

Shuoxi Zhang, Hanpeng Liu, and Kun He. 2024. Knowledge distillation via token-level relationship graph based on the big data technologies. *Big Data Research*, 36:100438.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.

Yuning Zhang, Buzhou Tang, Zhiwen Lu, Jianwu Yan, and Qing Xue. 2023. Codir: Contrastive distillation of intermediate representations for compressing pretrained transformers. In *Findings of ACL*.

Wei Zhou, Guoyin Zheng, Yanan He, Ting Yang, and Zhou Yu. 2022. Decoupled intermediate distillation. In *Findings of ACL*.