Scaling Down: Optimizing LLM Size for Specialized Domains with Knowledge Distillation

Aditya Ashvin
ashvin@usc.edu

Albert Kaloustian
kalousti@usc.edu

Pannawat Chauychoo
pchauych@usc.edu

Pooridon Rattanapairote
prattana@usc.edu

Abstract

This project investigates the feasibility of training smaller student models through knowledge distillation to achieve performance comparable to larger teacher models across domains such as mathematics, finance, history, and common sense reasoning. By leveraging a combined loss function that incorporates both student-teacher answer accuracy and mean squared error (MSE) between hidden layers, we demonstrate the potential of scaled-down LLMs for efficient deployment in resource-limited settings. Our findings show that student models can match or even exceed the accuracy of teacher models in specific tasks while significantly reducing inference time and computational costs, offering practical solutions for optimized model compression.

1 Introduction

1.1 Problem Significance

The rapid increase in large language model sizes (LLM) has resulted in great advancements in language understanding and generation, but these benefits have come at the expense of enormous computational and energy needs. The operation of large models in an environment constrained by computational resources is a common challenge. This challenge necessitates the use of smaller, more efficient models. Smaller LLMs rise to the occasion as a flexible solution which can be deployed at a relatively lower expense.

This project explores knowledge distillation as a method for compressing large, powerful language models into smaller, more efficient models capable of performing a variety of tasks. By transferring knowledge from a large teacher model to a smaller student model, we aim to retain performance while reducing the computational resources demanded to operate and maintain a larger model.

Efficient model compression has broad implications for applications needing on-device language processing, from mobile devices to robotics. This approach could democratize access to advanced AI by enabling high-performance models to run on less powerful devices without significant hardware or cloud costs. Such efficiency could empower smaller companies, independent developers, and new AI applications (focus on a particular domain) on devices like smartphones and smart glasses, expanding the accessibility and real-world impact of advanced AI technology.

1.2 Key Challenges

Key challenges include:

A big challenge we faced related to complex reasoning tasks. Logical reasoning tasks, such as Math, often require multiple steps, making it difficult for smaller models to perform effectively. Models need a substantial number of parameters to handle such reasoning tasks adequately Plaat et al. [2024]. Another challenge we faced was regarding memory constraints. The larger teacher model can cause out-of-memory (OOM) issues during training, necessitating the use of techniques such as gradient accumulation, custom validation, and mixed-precision training to optimize memory usage.

Computational resources were another key challenge we needed to overcome. Larger LLMs require training on systems with considerably larger CPU/GPUs available. Access to such resources can be scarce. Finally, we needed to find a way to deal with what we're calling 'Shortcutting'. When a small LLM initially faces a daunting challenge, its natural response is to shortcut its way to a smaller loss penalty. For instance, when presented with multiple-choice datasets, a small LLM might notice that one of the 4 options is the most common answer in the dataset, and choose only to give that answer moving forward to minimize the loss penalty. This kind of behavior prompts the integration of penalties specifically designed to discourage repetitive answers and reinforce diverse responses.

1.3 Beneficiaries

Edge environments, such as IoT devices, autonomous vehicles, and wearable health monitors, have limited processing power, storage, and battery life, making it challenging to install large-scale AI models. Smaller LLMs, trained by knowledge distillation from bigger models, can provide extremely effective, real-time responses while requiring fewer resources. Smaller LLMs can also be utilized in healthcare to provide real-time diagnostics and decision support for medical devices at the point of treatment.

Numerous industries and researchers seeking efficient AI solutions could also benefit by the less stringent resource requirements provided by small LLMs that could deliver comparable results to larger models. This would ultimately enable them to compete in an increasingly AI-driven landscape, while operating in a potentially resource-constrained environment.

2 Related Work

2.1 Existing Approaches

The research is primarily based on the approach outlined in Tian et al. [2024]. The paper discusses that smaller LLMs are more flexible and can be deployed at a lower expense as compared to larger models. A new knowledge distillation paradigm is introduced in which multiple larger teacher models train a smaller student model. The student model can inherit a broader range of skills and knowledge which leads to better generalization. The cross-entropy loss between the predicted output of the student and the teacher's output was employed in this research work. TinyLLM also demonstrates its superiority when compared to fine-tuning the whole dataset. Furthermore, the evaluation included six datasets spanning two reasoning tasks. A similar approach is taken in our research work where we explore the math, finance, history, and common reasoning domains to evaluate the performance of our student model (T5 small) on these domains when trained by larger teacher models. Mirzadeh et al. [2019] demonstrates that when there is a substantial gap between students and teachers, the performance of the student network deteriorates. Given a fixed student network, one cannot use an arbitrarily large student network which also means that a teacher can successfully impart knowledge to students up to a particular size and not less than this size.

Sanh et al. [2020] introduces a similar approach. The knowledge distillation leveraged in this research work shows that it is possible to reduce the size of a BERT model by 40% while retaining 97% of its capabilities and being 60% faster. The training objective involves minimizing the cross entropy between the predicted distribution of the model and the one-hot empirical distribution of the training labels. The research further demonstrates that the common dimensionality between teacher and student networks can be used as an advantage to initialize the student from the teacher by taking one layer out of two. Evaluation is done on the GLUE benchmark Wang et al. [2019]. A new two-stage learning framework is introduced in Jiao et al. [2020]. The two-stage learning framework includes

general distillation and task-specific distillation. A two-stage learning process is employed in our research work on the common sense reasoning domain.

Patient knowledge distillation for BERT compression is introduced in Sun et al. [2019]. The student model patiently learns from multiple intermediate layers of the teacher model following two strategies: 1. PKD-last which means learning from the last k layers and 2. PKD-skip which learns from every k layer. The research was evaluated on Sentiment classification, paraphrase similarity, Natural language inference, and much more. An important question is also answered in the research work, Does a better teacher help? The authors of the paper conclude that there is not much difference between the student's performance. In our research work, we have employed the T5 base as one of the teacher models based on the conclusion of Sun et al. [2019].Li et al. [2021] sees the distillation of knowledge in a fresh light, using the knowledge gap known as the residual, between a teacher and a student as guidance to train a much more lightweight student, called a res student. The authors claim that they achieve competitive performance with 18.04%, 23.14%, 53.59%, and 56.86% of the teachers' computational costs on the CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet datasets. In Wang et al. [2020], the student model is trained by deeply mimicking the self-attention module, which plays a pivotal role in transformer networks. Also, a teacher assistant is introduced in this work which helos the distillation of large pre-trained models. It is shown that 99% accuracy is retained on SQUAD 2.0 Rajpurkar et al. [2018] and other GLUE benchmarks using about 50% of the parameters. The paper also introduces task-specific distillation which is employed in our research work and data augmentation for downstream tasks.

Zhang et al. [2023] emphasize the benefits of multi-teacher knowledge distillation, where multiple pre-trained teacher models are used to supervise a single student model. This approach allows the student to leverage diverse knowledge sources from the teachers, enhancing its ability to generalize across tasks. By incorporating information from multiple domains or viewpoints, multi-teacher distillation provides richer supervision compared to traditional single-teacher methods. Our research adopts this multi-teacher strategy, utilizing FinBERT, RoBERTa, and BaseBERT as teacher models to train DistilBERT. This setup ensures the student inherits domain-specific expertise from FinBERT, linguistic capabilities from RoBERTa, and general knowledge from BaseBERT, resulting in a well-rounded and robust student model. Additionally, Luo et al. [2023] propose Deeply-Supervised Knowledge Distillation (DSKD), which focuses on aligning hidden layers between the teacher and the student using Mean Squared Error (MSE) loss to enhance the transfer of intermediate representations during training. Our research integrates MSE-based hidden layer alignment with the multi-teacher framework, ensuring consistent guidance across all layers of the student model while benefiting from the diverse knowledge provided by multiple teachers.

Recent research further explores advancements in knowledge distillation to improve efficiency and reduce computational demands. For instance, Juanhui Li [2024] proposes an active data selection mechanism that prioritizes high-quality, informative samples during the distillation process. This approach reduces the computational cost of training while maintaining high model performance. Additionally, Jongwoo Ko [2024] introduces a novel loss function based on skew Kullback-Leibler divergence and an adaptive off-policy optimization strategy, achieving significant reductions in distillation time while retaining accuracy. These methodologies highlight the potential for more resource-efficient and scalable distillation frameworks, aligning with our research objectives of improving inference efficiency and generalization in smaller student models.

To summarize, current approaches to knowledge distillation include:

- **Direct Knowledge Distillation:** Using cross-entropy loss between the teacher and student outputs, which allows for efficient training of student models but may lack the depth required for tasks requiring reasoning, such as common-sense understanding.

- **Layer Alignment Techniques:** Aligning intermediate layers between teacher and student models has shown promise for preserving model performance but can be computationally intensive and requires precise tuning.

- **Multi-Teacher Knowledge Distillation:** Utilizes multiple pre-trained teacher models to supervise a single student model, providing diverse knowledge from different domains, which enhances the student's generalization and robustness.

- **Deeply-Supervised Knowledge Distillation (DSKD):** Aligns intermediate hidden layers between teacher and student models using MSE loss, improving the transfer of intermediate representations and ensuring effective knowledge transfer throughout the training process.

2.2 Strengths and Weaknesses

The approach boasts several strengths and a few notable weaknesses. Key advantages include efficient mimicry, which accelerates the student model's learning by aligning its hidden layer representations with those of the teacher, though this compromises interpretability. Leveraging multiple teacher models allows the student to learn from diverse, complementary skills, enhancing adaptability and reducing overfitting to any single model. Pre-training the student model on common sense reasoning provides a foundational understanding, improving generalization and accelerating convergence during distillation. Additionally, the focus on multiple domains demonstrates the flexibility of knowledge distillation across various disciplines, addressing distinct challenges efficiently.

However, weaknesses remain. Interpretability issues arise from hidden layer alignment, making it difficult to understand the student model's decision-making process. Potential loss of fine-grained knowledge through cross-entropy loss might hinder reasoning tasks, though training on rationales and adding more teachers could mitigate this. The approach is also dependent on the quality of teacher models, as poor alignment can lead to less relevant knowledge transfer. Finally, the vague boundaries of size limitations in small LLMs create challenges in determining the optimal parameter count, with estimates varying widely across applications, adding some arbitrariness to model selection.

3 Datasets and Model Evaluation

3.1 Datasets

The evaluation of these methods primarily relies on accuracy metrics (Accuracy and F1 score).We are using multiple datasets to develop an extensive test set for evaluating our models across multiple domains.

Dataset Name	Number of Questions	Short Description
Winogrande	44,000	A pronoun disambiguation dataset with multiple choice questions, designed to test common sense reasoning.
Social IQA	33,000	Evaluates common sense reasoning by challenging models to grasp context and draw logical conclusions based on the presented information.
Trivia QA	650,000	Assesses the ability to extract and synthesize data from historical and trivia-related contexts.
Financial PhraseBank	4,840	Focuses on finance sentiment analysis with sentences from financial news articles annotated as positive, negative, or neutral, emphasizing the impact on stock prices.

Table 1: Datasets used for training and evaluation

- **Winogrande** Keisuke et al. [2019]: for common sense reasoning, a pronoun disambiguation dataset with multiple choice questions.
- **Social IQA** Maarten Sap [2019]: for common sense reasoning, evaluates how effectively the models grasp context and draw logical conclusions based on the information presented.
- **Trivia QA** Joshi et al. [2017]: for historical and trivia-related questions to assess the models' ability to extract and synthesize data from historical contexts.

- **Financial Phrasebank** Malo et al. [2014] for finance domain, a dataset with 4,840 sentences from financial news articles, each annotated for sentiment - positive, negative, or neutral, focusing on the potential impact of each sentence on stock prices.
- **AquaRat** Ling et al. [2017]: assesses mathematical reasoning by providing a rich variety of math-related questions that challenge models to solve issues using logical reasoning.

3.2 Model evaluation

We use the following models for teachers and student:

Domain	Student Model (parameters)	Teacher Model (parameters)
Common Sense	T5 Large (770M)	T5 XL (3B)
Trivia	T5 Small (60M)	T5 Large (770M)
Finance	DistilBERT (66M)	FinBERT (110M), BERT Base (110M), RoBERTa Base (125M)
Math	Flan T5 Small (80M)	Flan T5 Large (770M)

Table 2: Student and Teacher Models with Parameter Counts for Each Domain

Our project explores knowledge distillation across three domains: Common Sense Reasoning, Trivia, and Math. We validate model performance primarily using accuracy on validation data, applying loss functions tailored to each domain's specific needs. Each section below describes our approach and results achieved so far in these domains.

We have been measuring the student's accuracy against the teacher's on the domain specific validation sets on top of tracking the training loss. We will continue that to evaluate our approach and add check pointing to reserve the model's parameters and save progress. Our goal is to identify the optimal size for the small model to outperform the larger one in a domain specific area so checking at different sizes is required.

Our approach employs several key techniques. The project leverages several advanced techniques to optimize the training of a student language model with limited parameters. Knowledge distillation enables the student model to learn effectively from a teacher's outputs, enhancing performance despite its smaller size. Mixed-precision training reduces memory usage and accelerates computation, while gradient accumulation manages memory constraints by aggregating gradients over multiple batches before weight updates. The potential integration of DeepSpeed may further optimize training efficiency. The student model undergoes pre-training on an adjective disambiguation dataset to prepare for rationality-focused training, ensuring it is well-equipped for the task. Techniques like cross-validation and early stopping are used to fine-tune hyperparameters and prevent overfitting during training. Checkpointing ensures progress is consistently saved, enabling seamless resumption of training across sessions without loss of progress.

3.3 Common Sense Reasoning

The common sense reasoning (CR) portion of our project focuses on training a slightly larger LLM, T5-base (250M), as a student to learn from an even larger LLM, Flan-T5-XL (3B). With the expectation that the CR LLM would eventually need to tackle more language intensive tasks and performance metrics, it was decided that this model should be trained in a multi-stage approach.

- Stage 0: The pre-stage involved preprocessing and cross-validation to identify appropriate starting hyperparameters. 25 random sets of hyperparameters were each tested on 5% of the Winograd-XL dataset to determine which set performed best. The dataset was also preprocessed so that the model would be proposed with 2 options and would be expected to output a single integer as a choice between the options during training.
- Stage 1: In Stage 1, the student model was trained on the Winograd-XL dataset for pronoun disambiguation tasks. This stage served as foundational pre-training, enhancing the model's

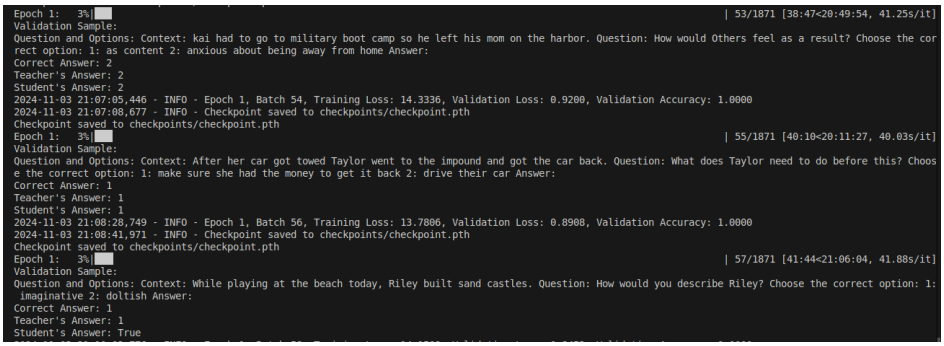

Figure 1: Early training examples of common sense reasoning teacher/student Stage 2 knowledge distillation training results.

capacity for language-intensive tasks and preparing it for knowledge distillation in the subsequent stages. The results from this stage demonstrated an improvement in the model's ability to handle pronoun resolution tasks, creating a solid base for further training.

- Stage 2: In Stage 2, the model was trained on the Social-IQA dataset for common sense reasoning tasks. A critical component of this stage was the use of a dual-loss approach, governed by an alpha value of 0.5. Half of the loss was calculated based on the cross-entropy difference between the student's predictions and the correct answers in the dataset. The other half utilized KL Divergence Loss to align the student's probability distributions with those of the teacher model. KL Divergence was specifically chosen because it encourages the student model to learn not only the correct answers but also the reasoning patterns and probabilistic outputs of the teacher, encouraging a deeper understanding of the task.

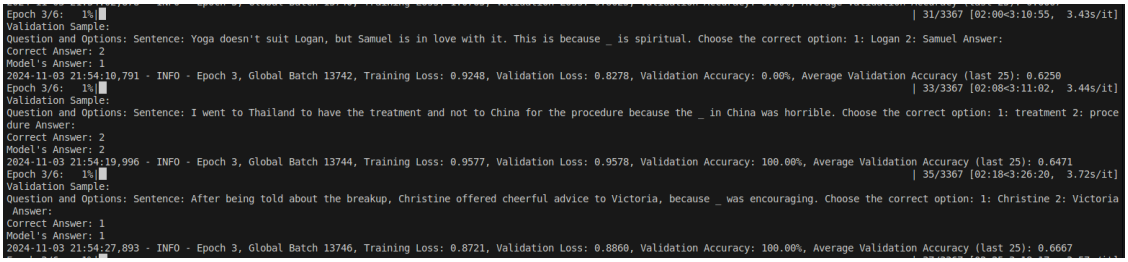

Figure 2: Examples of training on Winograd Pronoun Disambiguation.

3.4 Trivia

The Trivia QA dataset is used for the evaluation of the student and teacher models. Flan-T5 base(250M) and Flan-T5 small(77M) are the teacher and the student models used, respectively for the stage 1 training process. In the stage 2 training, the trained student model from stage 1 is used as the student model and the T5-large(770M) is the teacher model.

Training is based on a weighted loss, with alpha controlling the influence of the teacher's predictions. Key parameters reported per epoch include training loss, validation loss, and prediction accuracy for both teachers and students. After training, the student model generates predictions to evaluate its performance on previously unseen data. To measure the success of knowledge distillation, compare it with teacher forecasts and actual responses. This is the training and validation process. For testing, a set of unseen questions was fed to the models, and the output was matched to the given correct answers. The results are summarized in Table 4. To avoid overfitting, an early stop and learn rate scheduler was used. The student models aimed to achieve competitive performance while being optimized for computational efficiency.

Figure 3 shows the question, expected responses from student and teacher models, and the correct answer after training. In Question 1, the student model's prediction replicates the teacher's incorrect

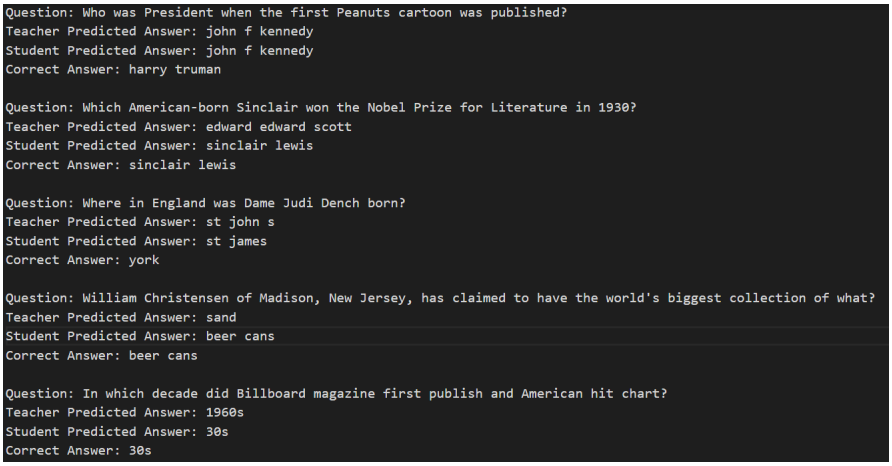

Figure 3: Some examples of the training results on Trivia QA dataset.

prediction, demonstrating the limitation of knowledge distillation when the teacher model is inaccurate: inaccuracies in the teacher's output might be passed down to the student model. Question 2, the student model accurately predicts the answer, demonstrating that it can learn from both the teacher's guidance and the actual labels. However, in Question 3, the student's answer differs from both the correct answer and the teacher's prediction, implying that, while knowledge distillation is useful, it can produce unexpected results if the student model fails to adequately handle the teacher's predictions with the ground truth.

3.5 Finance

The finance domain section focuses on leveraging sentiment analysis capabilities by training DistilBERT (66M) as the student model using three teacher models: FinBERT (110M), BaseBERT (110M), and RoBERTa (125M). Each teacher model contributes distinct domain-specific knowledge, with FinBERT specializing in financial sentiment, BaseBERT providing general-purpose language understanding, and RoBERTa offering linguistic refinement. The Financial Phrasebank dataset, comprising 4,840 labeled financial sentences, is utilized for this task.

Following the multi-teacher knowledge distillation strategy proposed by Zhang et al. [2023], we utilize outputs from multiple teacher models to enhance the diversity and robustness of the student model. Additionally, inspired by the Deeply-Supervised Knowledge Distillation (DSKD) framework introduced by Luo et al. [2023], we employ Mean Squared Error (MSE) loss to align intermediate representations between the student and teachers. This combination ensures that DistilBERT effectively captures the nuances of financial text, achieving a balance between compactness and performance.These loss functions ensure that DistilBERT effectively captures the nuances of financial text, achieving a balance between compactness and performance:

- **KL-Divergence Loss:** Measures the difference between the probability distributions of the student and teacher models' output logits, ensuring alignment in their predictions.
- **Mean Squared Error (MSE) Loss:** Aligns the hidden layer representations between the student and teacher models, enabling the student to capture intermediate-level knowledge.

Figure 4 shows predictions from the test set after full training. Each example includes the input sentence, the actual sentiment label, the student model's prediction, and predictions from the three teacher models (FinBERT, BaseBERT, and RoBERTa). It highlights cases where the student aligns with the teachers and areas where predictions differ, showing the impact of knowledge distillation and room for improvement.

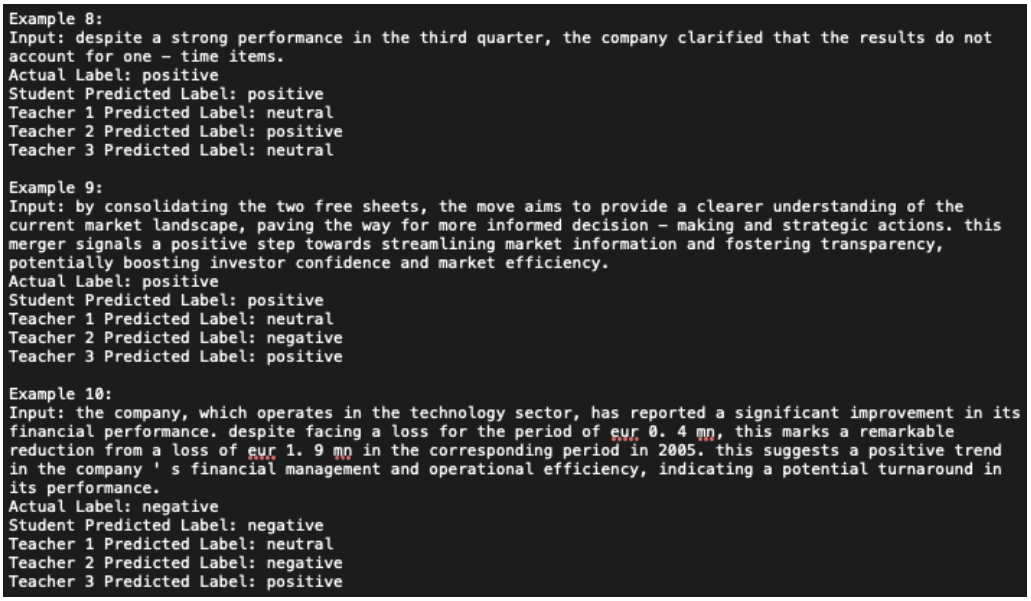

Figure 4: Some examples of the testing results on Financial Phrasebank Dataset

3.6 Math

In the Math domain, we applied knowledge distillation to improve the reasoning abilities of a student model on complex, multi-step problem-solving tasks using the AquaRat dataset. The training process combined two key components:

- **Answer Loss :** Compares the student's predicted answers with ground truth to refine its output predictions.
- **Hidden-Layer Loss :** Aligns the internal representations of the student model with the teacher model by projecting hidden states and minimizing the mean squared error (MSE) between them.

While the student model successfully outperformed the teacher model in accuracy in the AquaRat validation set (student: 27. 17%, teacher: 23.23%), the results highlight a limitation in achieving high accuracy on tasks requiring long chains of reasoning. Such tasks demand significantly larger model parameters to process extensive thought processes, which remains a challenge for lightweight student models. Consequently, we chose not to continue the experiment on math and instead focused on the other three tasks, which have lower depth in chain-of-thought reasoning and are more likely to yield clearer results.

Figure 5 shows the student model's performance on the test set after full training on AquaRat. Each includes a sample of math question proving that the student's predictions aligned with the teacher's one, having low accuracy on both

4 Results

4.1 Finance domain

Based on the results obtained from training and evaluating the DistilBERT student model using the Financial Phrasebank dataset, the overall findings demonstrate the effectiveness of the knowledge distillation approach. The student model successfully inherits key strengths from its teacher models (FinBERT, BaseBERT, and RoBERTa) and, in some cases, surpasses their individual performance. In particular, the integration of diverse teacher models enables the student to generalize better across financial sentiment classification tasks. Furthermore, the student model achieves this while maintaining high accuracy and F1 scores, demonstrating a balance between performance and efficiency.

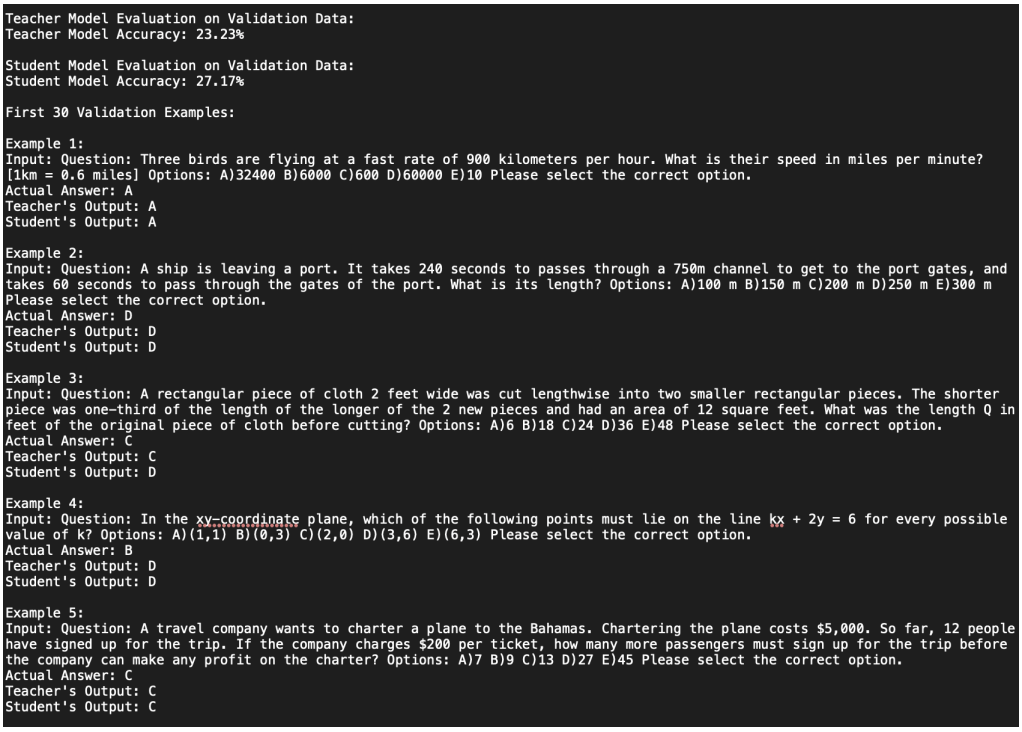

Figure 5: Some examples of the training results on AquaRAT

A particularly notable result is the student's ability to deliver predictions approximately twice as fast as its teacher models, demonstrating significant gains in inference speed without compromising accuracy. This efficiency highlights the potential for the use of the student model in real-time financial applications, where rapid decision-making is critical.

The graph in Figure 6 illustrates the reduction in training loss and validation loss over training steps. The training loss decreases steadily, reflecting the model's ability to learn effectively from the dataset. Meanwhile, the validation loss stabilizes around 0.5 after an initial decline, indicating that the model generalizes well to unseen data without significant overfitting.

Role	Model	Accuracy	F1 Score	Inference Time (ms)
Teacher	BERT Base	0.78	0.83	11.17
Teacher	FinPhrasebank RoBERTa	0.74	0.74	19.62
Teacher	FinBERT	0.86	0.84	10.21
Student	DistilBERT (Baseline)	0.33	0.32	5.19
Student	DistilBERT (Dataset Train)	0.78	0.78	5.19
Student	**DistilBERT (Distillation Train)**	**0.83**	**0.83**	**5.19**

Table 3: Performance comparison of teacher and student models in the finance domain, obtained using a GCP instance with an NVIDIA P100 GPU.

The table 3 compares the performance of teacher models (BERT, RoBERTa, FinBERT) and student models (DistilBERT variants) in terms of accuracy, F1 score, and inference time. The distilled DistilBERT model achieves a competitive accuracy of 83% and an F1 score of 0.83, outperforming some teachers like RoBERTa in accuracy. It also maintains a significantly lower inference time (5.96 ms), demonstrating efficient performance while preserving accuracy. The bolded row highlights the success of the distillation process in creating a compact yet effective student model. This experiment was conducted following the setup described by Sanh et al. Araci [2019], emphasizing the importance of training lightweight models while retaining the performance of larger teacher models.

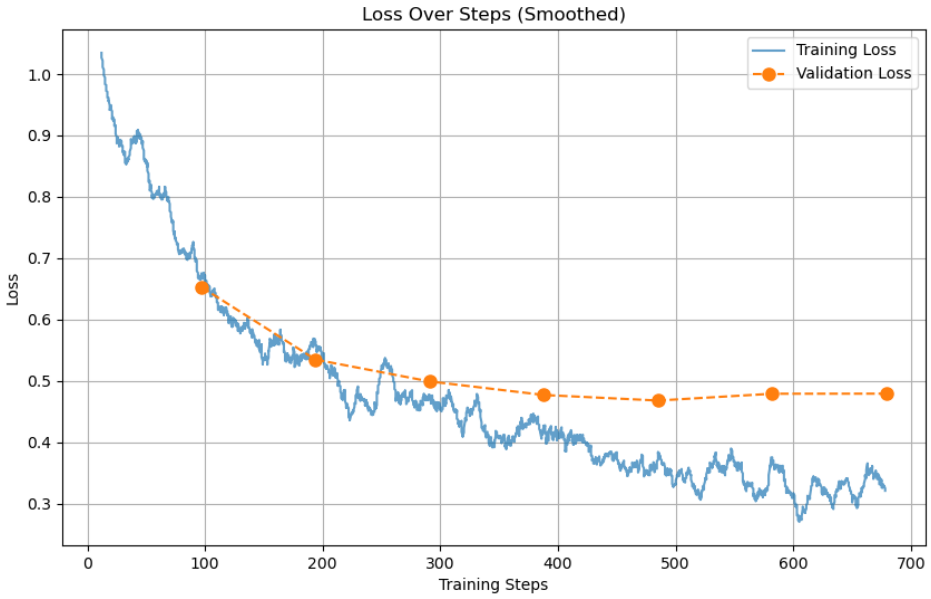

Figure 6: Training and Validation Loss for the Finance domain

4.2 Trivia domain:

While training the student models we saw consistent improvement across 100 epochs with the student achieving 95% of the teacher's accuracy, albeit on the lower side at 45% overall. The graph in Figure 7 shows a sharp drop in training and validation loss within the first few epochs, quickly stabilizing at a low level. This suggests efficient learning of factual content, with stable losses across epochs indicating good generalization in the history trivia domain.

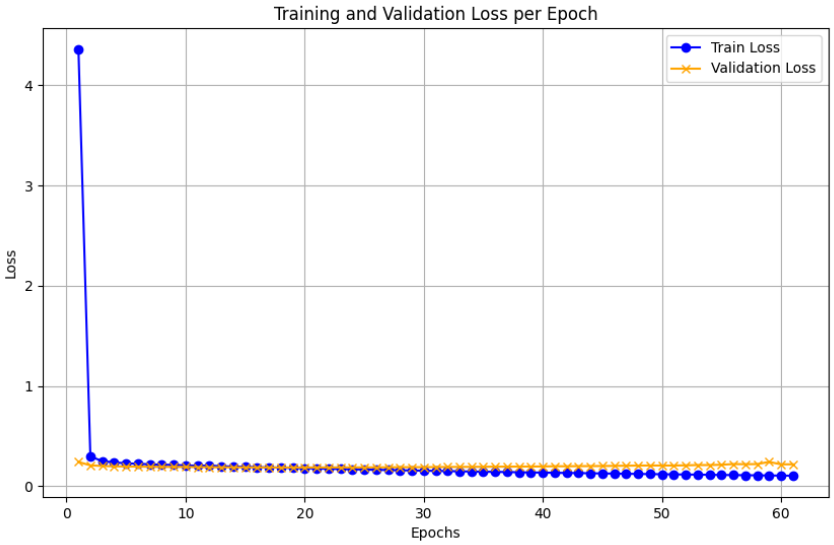

Figure 7: Training and Validation Loss curves for the trivia domain.

Table 4 compares the F1 score, precision, recall, and number of questions answered per second by the teacher and the student models. The table shows the effectiveness of knowledge distillation from teacher to student models while improving computation efficiency. The T5 small model which is our baseline model achieved an F1-score of 0.6, a precision of 0.48, and a recall value of 0.81. As expected, the T5 small achieves lower performance than the T5 base which has a precision value of 0.68, recall of 0.88 and a F1-score of 0.76 and the T5 large which have a precision value of 0.82, recall of 0.85 and F1 score of 0.84. However, these larger teacher models take a lot of time to process the answers as shown in the number of questions answers per second column. T5 small answers about 10.79 whereas the T5 base and T5 large have values of 10.48 and 5.14 respectively.

Model Name	F1 Score	Precision	Recall	No. of Questions/Sec
T5 Small (Baseline)	0.60	0.48	0.81	10.79
T5 Base (Teacher 1)	0.76	0.68	0.88	10.48
T5 Large (Teacher 2)	**0.84**	**0.82**	0.85	5.14
T5 small (Stage 1 Student)	**0.84**	0.76	**0.93**	**14.98**
T5 small (Stage 2 Student)	**0.84**	0.76	0.92	14.96

Table 4: **Trivia Domain:** F1 score, Precision, Recall, and the number of questions per second comparison of the baseline model, teacher models, and the trained stage 1 and stage 2 student models.

The trained student models (stage 1 and stage 2) make a significant progress in striking a balance between accuracy and efficiency. The F1 scores of the trained student model match the T5 large model and outperforms both the T5 small baseline model and the T5-base model by 40% and 10.5% respectively. The student models also have improved recall values of 0.93 and 0.92 for the stage 1 and the stage 2 models respectively. The number of questions answered per seconds values of the trained student models also outperform the baseline and the teacher models by 32%.

The two-stage distillation shows marginal improvement particularly in the recall and the F1-score. Through stage 1 training, the student model is already able to achieve performance parity with the larger T5 Large teacher by distilling a significant part of the teacher model's knowledge. By now, much of the learning potential of the model has been efficiently used. In the second stage, more distillation occurs, which mostly refines the student model and results in little changes rather than notable performance improvements. As per Kim et al. [2022], the results we got in our research is consistent to the idea that the goal of knowledge distillation is to strike a balance between the efficiency and accuracy.

4.3 Common Sense Reasoning domain:

The training loss and validation loss for pronoun disambiguation demonstrated a sharp decline in the initial stages, stabilizing below 0.5, which indicates efficient learning and minimal overfitting. Pronoun disambiguation accuracy consistently improved throughout training, showing the model's capacity to generalize from the teacher's outputs.

For stage 2, the validation loss improved steadily until roughly 10,000 samples of training in, at which point training loss continued to improve while validation loss flattened and began to reverse, as shown in figure 10. With the aid of our checkpointing strategy which routinely creates checkpoints every specified number of training samples, we were able to revert the model to the optimal point in the model's training and early-stop the training there.

Model Name	Accuracy	No. of Seconds/Question
T5 XL (Teacher)	0.81	12.5
T5 Large (Stage 1 Student)	0.72	5.1
T5 Large (Stage 2 Student)	**0.79**	**5.3**

Table 5: **Common Sense Reasoning:** Accuracy and number of seconds per question comparison of the teacher model and the trained stage 1 and stage 2 student models.

The table 5 summarizes the accuracy and number of seconds per question after the two-stage training. Though the accuracy still trailed the teacher model, the significantly improved inference time for

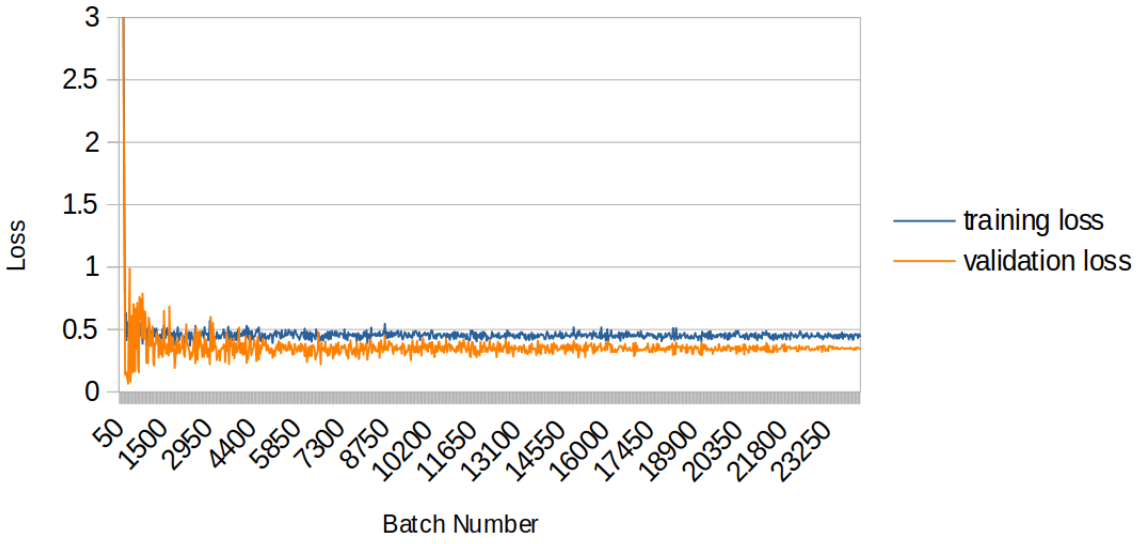

Figure 8: Loss after 23000 training samples for pronoun disambiguation

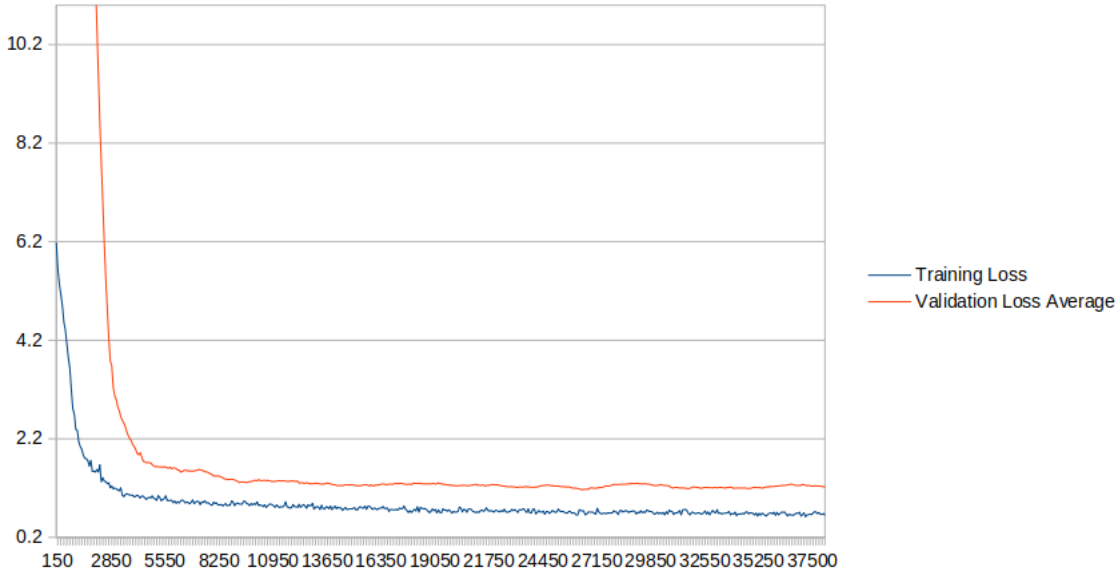

Figure 9: Common sense reasoning training and validation loss, with a rolling average to smooth noise, after 37000 training samples in stage 2 schema

propagating answers to common sense reasoning questions is worthy of note. This highlights the models performance which nearly rivals the teacher model at a significantly faster pace.

5 Conclusion and Future Scope

5.1 Conclusions from Experiments

Our exploration into optimizing LLMs for specific tasks through knowledge distillation has demonstrated promising results, particularly for constrained environments. By training smaller student models to emulate larger teachers, we observed that student models could achieve comparable, and sometimes even superior performance in targeted domains such as common sense reasoning, history, and finance, while requiring fewer computational resources.

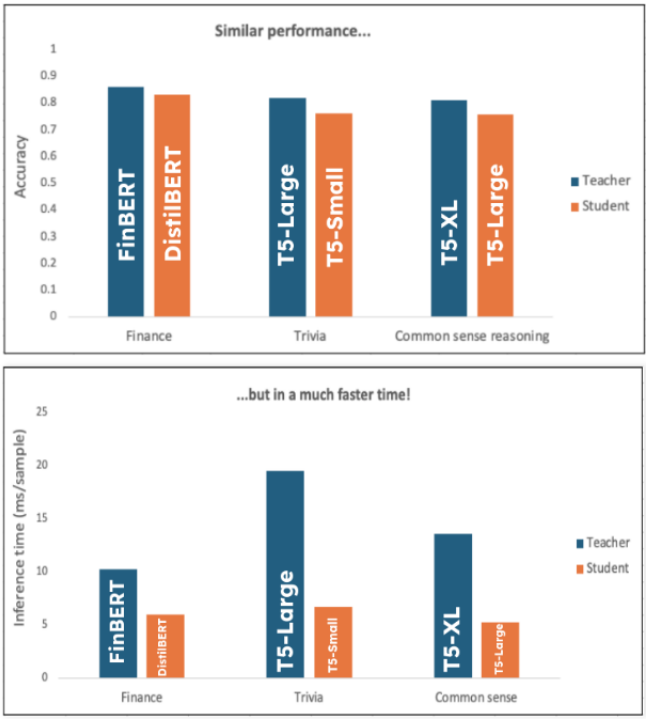

Figure 10: Accuracy and inference time comparisons for 3 domains

In figure 10, it highlight that student models achieve similar accuracy to their teacher models across all domains, including Finance, Trivia, and Common Sense reasoning, demonstrating the effectiveness of the knowledge distillation process. However, student models significantly outperform teacher models in inference speed, requiring substantially less time per sample, making them more efficient for real-time applications.

However, our experiments revealed several unique challenges. First, while student models can quickly learn from teacher's outputs, they often exploited inherent biases in the dataset or took shortcuts like only giving the most common answers, compromising their test performance in logical tasks. To combat this, we implemented custom loss functions and penalties like KL Divergence loss to discourage repetitive, superficial solutions and force student to learn task-specific nuances. Although we did see improvements, we decided to pivot away from heavy reasoning tasks like math since our smaller student models were not able to develop the emergent capabilities of the larger ones. Another major insight from our work is the critical importance of standardizing output structures across training stages. Models trained on different answer formats struggled to adapt when transitioning between datasets with varying numbers of choices. Implementing consistent answer formats (e.g., integers representing options) across datasets significantly stabilized the model's learning process.

Nevertheless, there were surprising advantages with our approach. Firstly, we found that students learn better from multiple teachers and can develop an exponential increase in inference speed while maintaining superior accuracy. Secondly, processioning data to ensure proper student output based on the dataset minimizes training time and promotes consistent learning. Thirdly, two-stage knowledge distillation balances the efficiency and accuracy of the model and prevents overfitting. Lastly, we utilized many tailored loss functions across our domain-specific tasks and saw noticeable increase in performance overall, suggesting that domain specific tasks require their own loss functions.

5.2 Future Implementation

We plan to explore fine-tuning loss functions and regularization methods further to counteract the model's tendency to rely on trivial solutions. Our findings suggest that smaller models can indeed inherit substantial reasoning capabilities from their larger counterparts, provided training strategies effectively mitigate shortcut behaviors. Ultimately, we envision this approach enabling compact,

efficient models suited for deployment in resource-limited environments without compromising on task performance.

References

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020. URL https://arxiv.org/abs/1909.10351.

Tianyi Chen Se-Young Yun Jongwoo Ko, Sungnyun Kim. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Hui Liu Xianfeng Tang Sheikh Sarwar Limeng Cui Hansu Gu Suhang Wang Qi He Jiliang Tang Juanhui Li, Sreyashi Nag. Learning with less: Knowledge distillation from large language models via unlabeled data, 2024. URL https://arxiv.org/abs/2411.08028.

Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://huggingface.co/datasets/allenai/winogrande.

Minsoo Kim, Sihwa Lee, Sukjin Hong, Du-Seong Chang, and Jungwook Choi. Understanding and improving knowledge distillation for quantization-aware training of large transformer encoders, 2022. URL https://arxiv.org/abs/2211.11014.

Xuewei Li, Songyuan Li, Bourahla Omar, Fei Wu, and Xi Li. Reskd: Residual-guided knowledge distillation. *IEEE Transactions on Image Processing*, 30:4735–4746, 2021. ISSN 1941-0042. doi: 10.1109/tip.2021.3066051. URL http://dx.doi.org/10.1109/TIP.2021.3066051.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

Shiya Luo, Defang Chen, and Can Wang. Knowledge distillation with deep supervision, 2023. URL https://arxiv.org/pdf/2202.07846.

Derek Chen Ronan LeBras Yejin Choi Maarten Sap, Hannah Rashkin. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65, 2014.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. URL https://arxiv.org/abs/1902.03393.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey, 2024. URL https://arxiv.org/abs/2407.11511.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. URL https://arxiv.org/abs/1806.03822.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019. URL https://arxiv.org/abs/1908.09355.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V. Chawla. Tinyllm: Learning a small student from multiple large language models, 2024. URL https://arxiv.org/abs/2402.04616.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Hailin Zhang, Defang Chen, and Can Wang. Adaptive multi-teacher knowledge distillation with meta-learning, 2023. URL https://arxiv.org/pdf/2306.06634.