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 a smaller student model through knowledge distillation, focusing on achieving performance comparable to a larger teacher model. We aim to determine the minimum number of parameters required for the student model to succeed in various domains such as math, finance, history, and commonsense reasoning. The approach leverages a combined loss function that incorporates both student-teacher answer accuracy and mean squared error (MSE) between hidden layers, offering insights into efficient model training. Our results highlight the potential of scaled-down LLMs for efficient, accurate deployment in resource-limited settings, offering valuable insights into optimized model compression.

1 Introduction

1.1 Problem Significance

The rapid rise of large language model (LLM) sizes has resulted in great advancements in language understanding and generation, but these benefits have come at the expense of enormous computational and energy needs. Operating 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:

- **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].
- **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:** Larger LLMs require training on systems with considerably larger CPU/GPUs available. Access to such resources can be scarce.
- **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: Reiterate the current approaches

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.

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 T5-base as one of the teacher models based on the conclusion of Sun et al. [2019]. 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.

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.

2.2 Strengths and Weaknesses

The strengths of our approach include:

- **Efficient Mimicry:** This technique aligns the student's hidden layer representations to the ones of the teacher, allowing the student to learn more complex patterns and representations faster. However, focussing on hidden layer alignment has a drawback: it limits interpretability.
- **Multiple Teacher Models:** By learning from teachers with complementary skills, the student encounters a more robust learning process that eliminates overfitting to the constraints of any particular model, ideally resulting in improved accuracy and adaptability across a broader range of settings.
- **Pre-training student model:** By pre-training the student model on commonsense reasoning before engaging in teacher-student training, our approach aims to provide the student model with a foundational understanding that could enhance its ability to align with the teacher's outputs. This preparatory stage may improve the student's capacity to generalize and accelerate convergence during the distillation process, potentially yielding better results.
- **Multiple Domains:** One of our focuses in this project was how multiple different disciplines could benefit by using the knowledge distillation method. Each field has distinct obstacles and learning possibilities, which can be efficiently addressed by targeted knowledge distillation.

However, weaknesses persist:

- **Interpretability Issues:** When training emphasises on hidden layer alignment, it is frequently unclear why the student model makes certain decisions or which features it prioritises at different stages. The student model learns to duplicate patterns without having a clear concept of what each hidden state represents or why certain transformations are ideal.
- **Potential Loss of Fine-Grained Knowledge**: Although cross-entropy loss is effective in aligning the student model's predictions with those of the teacher models, it may not capture the finer, more complex knowledge representations needed for reasoning tasks like common sense reasoning. We aim to address this by training on rationales and adding more teachers.
- **Dependency on Teacher Model Quality**: If the teacher models are not well-aligned with the student's domain tasks, the student model may inherit knowledge that's less relevant or beneficial for the intended applications.

- **Vague Limitation Boundaries:** Emergent behavior such as advanced logic and chain-of-thought capabilities becomes available when a model becomes large enough, but knowing how large a model needs to be can be difficult to pinpoint. Various papers give different estimations based on different fields of application, so choosing the size of small LLMs to test can seem somewhat arbitrary. The best resources available in this domain are those that have experimented previously in the small LLM field.

2.3 Evaluation Methods

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.

- **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.
- **Trivia QA** Joshi et al. [2017]: for historical and trivia-related questions allows us to assess the models' ability to extract and synthesize data from historical contexts.
- **Winograd** Keisuke et al. [2019]: for common sense reasoning, where we can evaluate how effectively the models grasp context and draw logical conclusions based on the information presented.

3 Methods: What is your plan to approach it? What is the progress you have made?

3.1 Methods Employed

We use the following models for teachers and student:

- Teachers: Flan T5 base (250M), Flan T5-XL (3B)
- Student: Flan T5 small (77M), Flan T5 base (250M)
- Tasks: Math, Trivia QA, Commonsense reasoning

Our approach employs several key techniques:

- **Knowledge Distillation:** This technique allows the student model to learn from the teacher's outputs, facilitating improved performance despite having fewer parameters.
- **Mixed-Precision Training:** This technique reduces memory usage and speeds up computation, crucial for training on limited hardware.
- **Gradient Accumulation:** This method helps manage memory constraints by accumulating gradients over multiple batches before updating the model weights.
- **DeepSpeed Integration:** Potential use of DeepSpeed to further optimize training efficiency.
- **Pre-training student model:** Before teacher/student training commences, the commonsense reasoning student LLM undergoes pre-training on a adjective disambiguation dataset to prepare for upcoming rationality training.
- **Cross-validation:** Identifying optimal hyperparameters prior to training ensures optimal learning progress.
- **Early stopping:** By utilizing periodic validation checks during training sessions, it's easier to check if training loss and validation loss are beginning to diverge to stop training early before over-fitting.
- **Checkpointing:** Training isn't always able to be performed in a consistent session, and having to restart from the beginning can slow progress. We've integrated regular checkpointing to save the model's training progress over time so progress can be loaded and continued, even if portions of a training session occur hours apart.

3.2 Current Progress

Our project explores knowledge distillation across three domains: Commonsense 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.

3.2.1 Commonsense Reasoning

The commonsense 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.

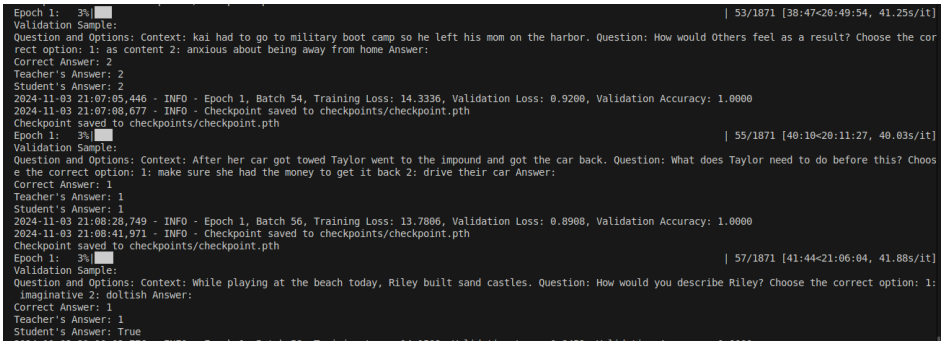

Figure 1: Early training examples of commonsense reasoning teacher/student Stage 2 knowledge distillation training results.

- 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: The model is trained on the Winograd-XL dataset on pronoun disambiguation. This pre-training step prepares the CR model for more language intensive tasks to follow, as well as for knowledge distillation with the teacher LLM. Results of Stage 1 training were roughly

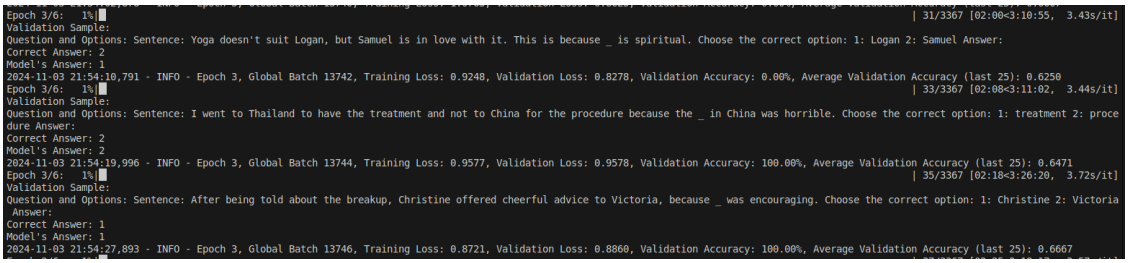

Figure 2: Examples of training on Winograd Pronoun Disambiguation.

3.2.2 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. 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 to teacher forecasts and actual replies.

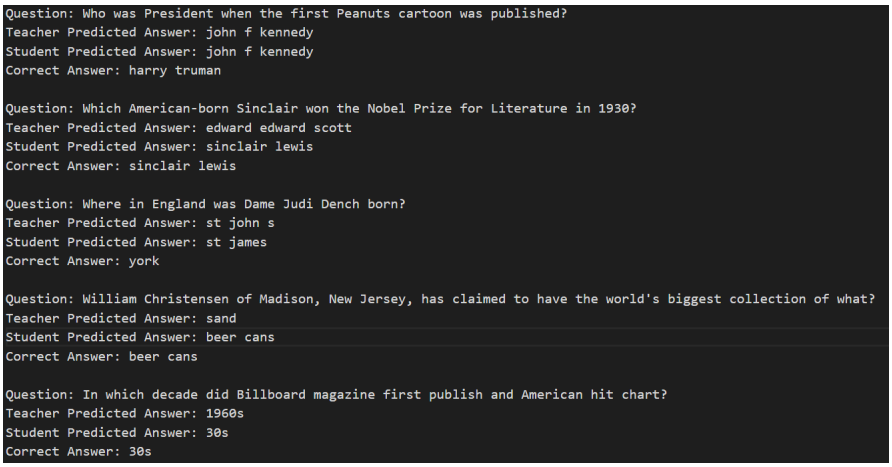

Figure 3: Some examples of the training results on Trivia QA dataset.

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 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.

3.3 Math

The Math portion of our project focuses on applying knowledge distillation to improve the student model's reasoning abilities on complex, multi-step problem-solving tasks using the AquaRat dataset. Our approach combines answer loss and hidden-layer loss, each contributing equally to the total loss:

- Answer Loss (50%): This component compares the student's predicted answer directly with the ground truth.
- Hidden-Layer Loss (50%): The hidden states from the teacher model are projected to match the hidden size of the student model, and the mean squared error (MSE) is calculated between these teacher and student hidden states. This loss component helps align the internal representations of the student with those of the teacher to enhance learning transfer.

Figure 4 showcases the student model's performance on the validation set after full training on AquaRat. Each includes a math or reasoning question, displaying the actual answer, teacher's prediction, and student's prediction. The results reflect the student model's accuracy and alignment with the teacher's predictions.

Key achievements in the Math domain include:

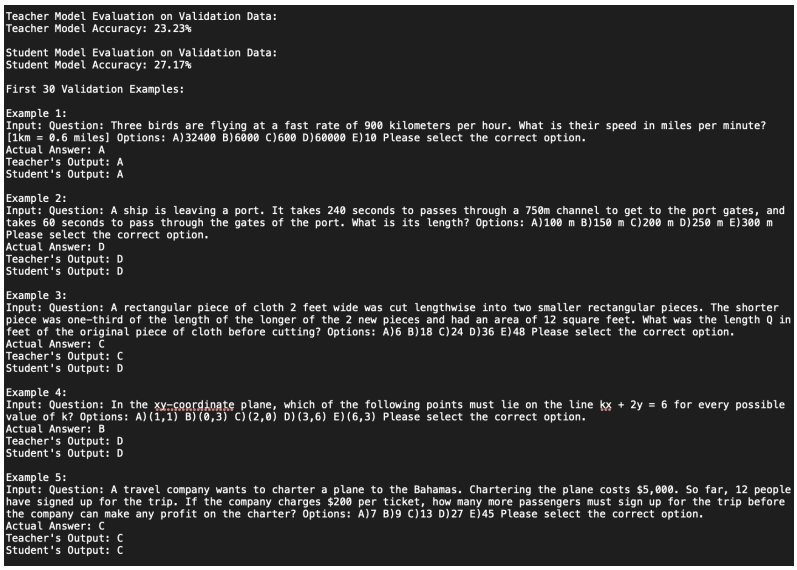

Figure 4: Some examples of the training results on AquaRAT

- Successfully trained the student model to outperform the teacher model in terms of accuracy by testing on validation set of AquaRat dataset: Teacher Accuracy: 23.23% vs. Student Accuracy: 27.17%.
- Demonstrated that hidden-layer loss contributes positively to model accuracy, as evidenced by slight improvements in performance.
- Completed training on the full AquaRat dataset without encountering out-of-memory (OOM) issues, affirming the stability of our approach on resource-constrained hardware.

3.4 Next Steps

Future plans involve:

- Integrating additional teacher models to assess the impact on student's performance.
- Further tuning hyperparameters to enhance model accuracy.
- Conducting generalization and normalization tests to ensure robustness across different datasets.
- Testing the student on other subjects like law, biology and finance.
- Creating more custom loss functions to promote actual learning instead of copying.
- Checking different student's sizes and their pre-trained tendencies.
- Stage 2 of the CR model training, where the student model is trained on a custom loss function comparing hidden layer logits to the teacher LLM as well as final output cross-entropy loss.

3.5 Updated Timeline

Here's the revised timeline for our project:

- Weeks 1-2 (Sept 23 - Oct 6): Conduct literature review and analyze TinyLLM framework.
- Weeks 3-4 (Oct 7 - Oct 20): Implement student-teacher framework, select models and fine-tuning methods.
- Weeks 5-6 (Oct 21 - Nov 3): Run experiments on chosen datasets.
- Weeks 7-8 (Nov 4 - Nov 17): Implement multiple teacher into framework & run more experiment on other datasets and domains

- Weeks 9-10 (Nov 18 - Dec 1): Finalize experiments, document findings, prepare final presentation and report.

4 Experiments: What are the results you have achieved so far? What are the results you are missing?

4.1 Current Results

Our experiments have shown minor improvements in performance when the student model mimics only one teacher. So far, we have observed:

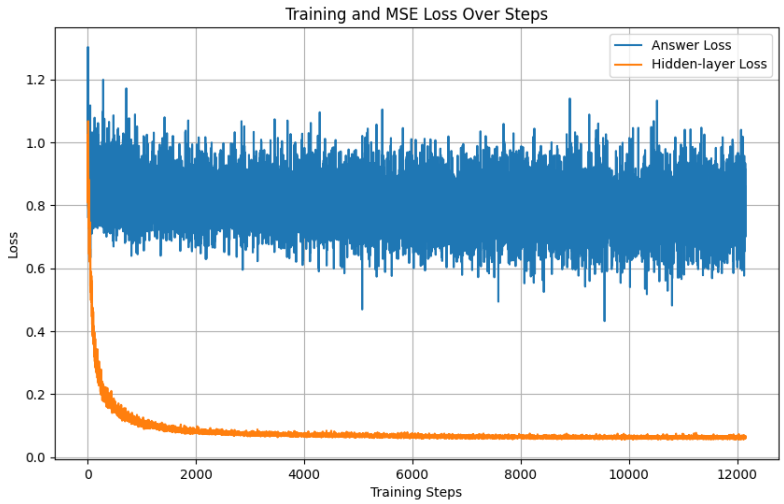

Figure 5: Training and MSE Loss for the Math domain

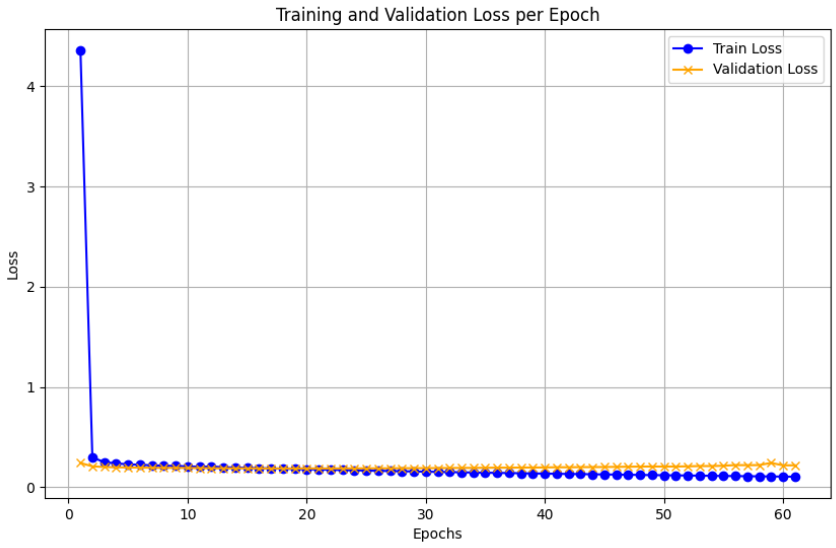

Figure 6: History Trivia - Loss after 10,000 training samples for Trivia QA.

- **Math domain:** During math training, the student model showed a significant decrease in answer loss within the first 2000 steps, dropping from over 1.2 to a steady range around 0.7–0.9. This early reduction reflects the model's rapid adaptation to the task, after which answer loss stabilizes, showing only minor fluctuations. Meanwhile, hidden-layer loss remained consistently low, indicating effective alignment with the teacher model's internal representations.

  The graph in Figure 5 illustrates these trends, with answer loss sharply decreasing during the initial steps and then settling into a stable range. Hidden-layer loss stays low throughout the training process, emphasizing stable knowledge transfer from the teacher to the student model
- **Trivia (History) domain:** Consistent improvement across 100 epochs with the student achieving 95% of the teacher's accuracy, albeit on the lower side at 45% overall. Looking to add Flan T5 large as a teacher and test if the student can generalize since it is all fact-based.

  The graph in Figure 6 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.

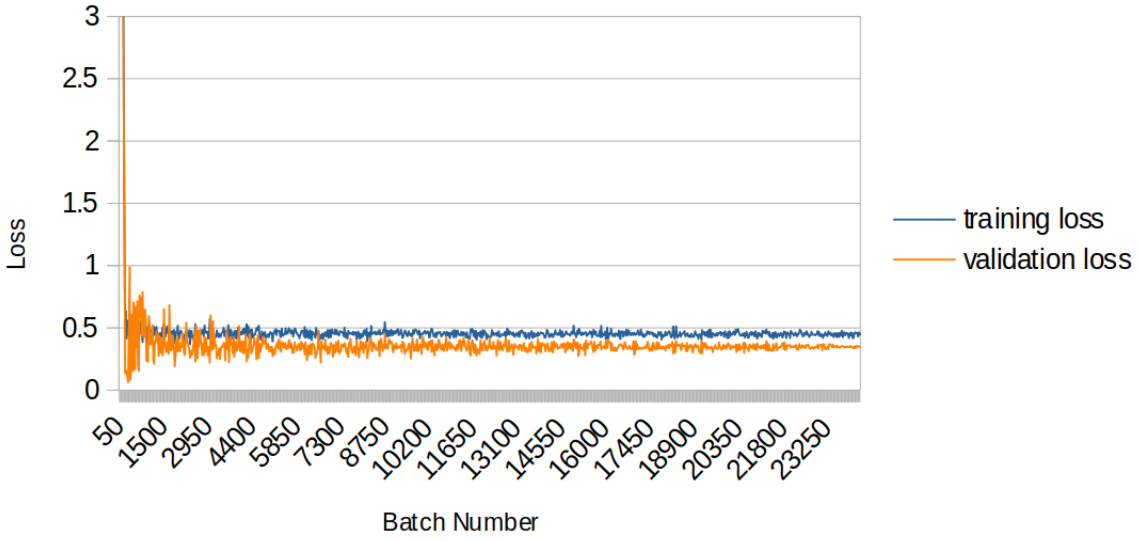

Figure 7: Loss after 44000 training samples for pronoun disambiguation

- **Commonsense Reasoning domain:** Preprocessed dataset to ensure proper student LLM output and used cross-validation to determine optimal hypermeters. Performed Stage 1 Training of student LLM on Winograd-XL dataset for pronoun disambiguation for approximately 6 hours, spanning 5 epochs. Results are within expectations prior to shifting into Stage 2 where knowledge distillation will continue on the SocialIQA dataset. **?**

  The graph in 7 shows an initial sharp drop in both training and validation loss, quickly stabilizing below 0.5. This indicates efficient learning and stable performance on pronoun disambiguation. The close alignment between training and validation loss suggests minimal overfitting, highlighting consistent generalization throughout Stage 1.

4.2 Missing Results

While our project has demonstrated initial success in training student models that approximate or even exceed the performance of teacher models in certain domains, several key results are still needed to fully evaluate our approach. The following areas require further experimentation and data collection:

- **Per-Task Error Analysis:** Given that our datasets span multiple domains, including math, commonsense reasoning, and history, it is essential to perform an in-depth error analysis by task. By evaluating where the model's predictions deviate from the teacher's responses

or ground truths, we can identify task-specific weaknesses, such as a tendency to oversimplify math problems or misinterpret historical context. This analysis would inform future adjustments in model architecture and distillation strategy to better handle these nuanced challenges.

- **Multi-teacher distillation results:** Our current approach utilizes a single teacher model per task. To validate our hypothesis that multiple teachers can provide a more comprehensive learning experience, we plan to incorporate several teacher models with complementary strengths. We need results comparing single-teacher and multi-teacher distillation to determine whether the added complexity yields significant performance gains for the student model.
- **Performance on reasoning-intensive tasks:** As logical reasoning and multi-step problem-solving tasks (e.g., advanced math) remain challenging for smaller models, additional evaluations are needed to better understand the limitations. Testing the model on complex reasoning tasks with varying parameter sizes will help us assess the "ceiling" of performance for smaller models and identify tasks that may inherently require larger models.
- **Generalization across domains:** We need additional experiments with datasets from diverse fields, such as finance, to determine whether our knowledge distillation techniques are universally effective or if they need to be specialized for specific domains. This will help us evaluate the robustness of our model across tasks requiring different levels of reasoning and contextual understanding.

5 Discussion: What can you learn from your results? What are the next steps for the project?

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 commonsense reasoning, history, and basic math, while requiring fewer computational resources.

However, our experiments revealed several unique challenges. First, while student models can quickly learn from teacher outputs, they often exploit dataset biases or shortcut patterns, compromising their performance in genuine reasoning tasks. For instance, in multiple-choice formats, models often defaulted to the most frequent answer rather than engaging in meaningful learning. This observation underscores the necessity for custom loss functions and penalties that discourage repetitive, superficial solutions, reinforcing deeper learning of task-specific nuances.

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.

Future work will expand our training framework to incorporate multiple teacher models, aiming to distill diverse knowledge into a single student. Additionally, we plan to explore fine-tuning loss functions and regularization methods further to counteract the model's tendency to rely on trivial solutions. Our initial 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.

5.2 Expectation vs. Reality

These results align with our expectations, given the inherent challenges associated with resource constrained LLMs on a variety of task types. While the initial performance of the students have been promising, we recognize the need for further enhancements and will continue to expand upon our ideas.

5.3 Future Implementation

The next steps involve:

- **Multiple teacher knowledge distillation frameworks:** To improve the effectiveness of knowledge transfer, we plan to integrate additional teacher models in our framework. We hypothesize that by training a single student LLM from multiple teach LLMs, the student might pick up some characteristics from each which might allow it to perform better than either teacher in a given field.
- **Experimenting with diverse datasets:** We will broaden the training scope by testing on datasets in finance, where the chain of reasoning tends to be shorter and less complex than in math, potentially improving efficiency. This will help determine whether knowledge distillation in fields with simpler reasoning demands can achieve accuracy more quickly and effectively.
- **Cross-validation to tune hyperparameters:** Effective tuning will involve both standard training parameters—such as learning rate, batch size, dropout rate, weight decay, and gradient-accumulate steps and distillation-specific hyperparameters, like the weight assigned to semantic loss in our two-stage distillation approach and the hidden-layer alignment weight in the hidden-layer approach. Fine-tuning these weights will allow us to balance the influence of the teacher's knowledge on the student, optimizing for the best performance without overloading the student with unnecessary complexity. We may also explore dynamically adjusting these weights during training to enhance adaptability across different learning stages.
- **Investigating the limits of model accuracy on advanced datasets:** As we aim to determine how compact a student model can be while maintaining accuracy levels close to those of larger teacher models, there may be inherent limitations in applying our approach to domains like math, especially with complex, multi-step reasoning tasks such as those in AquaRat. These tasks often require extensive chain-of-thought processing, which could push our current framework to its upper limit of effectiveness and reveal a potential "ceiling effect" where further miniaturization leads to unacceptable drops in accuracy. In these cases, we'll determine whether increasing the model size is necessary or whether further model optimization will be possible.

References

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.

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.

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.

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

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.