# Knowledge Distillation of LLMs for Automatic Scoring of Science Assessments

Ehsan Latif[1,2], Luyang Fang[1,3], Ping Ma[3,*], and Xiaoming Zhai[1,2,*]

[1] AI4STEM Education Center, Athens, GA, USA
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA
[3] Department of Statistics, University of Georgia, Athens, GA, USA
*{pingma,xiaoming.zhai}@uga.edu

**Abstract.** This study proposes a method for knowledge distillation (KD) of fine-tuned Large Language Models (LLMs) into smaller, more efficient, and accurate neural networks. We specifically target the challenge of deploying these models on resource-constrained devices. Our methodology involves training the smaller student model (Neural Network) using the prediction probabilities (as soft labels) of the LLM, which serves as a teacher model. This is achieved through a specialized loss function tailored to learn from the LLM's output probabilities, ensuring that the student model closely mimics the teacher's performance. To validate the performance of the KD approach, we utilized a large dataset, 7T, containing 6,684 student-written responses to science questions and three mathematical reasoning datasets with student-written responses graded by human experts. We compared accuracy with state-of-the-art (SOTA) distilled models, TinyBERT, and artificial neural network (ANN) models. Results have shown that the KD approach has 3% and 2% higher scoring accuracy than ANN and TinyBERT, respectively, and comparable accuracy to the teacher model. Furthermore, the student model size is 0.03M, 4,000 times smaller in parameters and x10 faster in inferencing than the teacher model and TinyBERT, respectively. The significance of this research lies in its potential to make advanced AI technologies accessible in typical educational settings, particularly for automatic scoring.

**Keywords:** large language model (LLM) · BERT · knowledge distillation · automatic scoring · education technology

## 1 Introduction

Artificial Intelligence (AI) in education has evolved from a theoretical concept to a practical tool, significantly impacting classroom assessment practices and adaptive learning systems [4,6]. AI for personalized learning and assessment provides opportunities for more tailored and effective educational experiences [13]. Integrating Large Language Models (LLMs) from domains on AI like BERT [1] into education has been a significant milestone in enhancing learning experiences, providing personalized learning content and support, and facilitating automatic

scoring [9,11,10,14]. Despite their potential, the deployment of LLMs in educational settings is constrained by their considerable size (714MB for 178 million parameters and 495MB for 124 million parameters) and computational requirements (16 Tensor Processing Units), presenting a challenge for widespread adoption in resource-constrained educational environments such as mobiles/tablets and school-provided laptops with no GPUs or TPUs and limited memory [5].

To bridge this gap, our study explores the feasibility of distilling the knowledge of LLMs into smaller neural networks, referred to as *student models*, with fewer parameters and hidden layers. By training a smaller student model using soft labels provided by a fine-tuned LLM (i.e., *teacher model*), we aim to achieve a similar scoring performance to that of LLMs, but with reduced model size.

The significance of this research lies in its potential to make advanced AI technologies accessible in typical educational settings. The study addresses the technical challenges of deploying AI models in resource-constrained environments and highlights the potential of AI to transform educational assessment practices. By enabling the deployment of efficient automatic scoring systems on less powerful hardware available in school settings, we contribute to the democratization of AI in education. The key contributions of this paper are:

- We demonstrate the successful application of a novel knowledge distillation (KD) strategy that, while inspired by [5], is uniquely adapted and optimized for the context of educational content.
- Our approach achieves a significant reduction in model size and computational requirements without compromising accuracy. The student model, distilled from a fine-tuned BERT teacher model, exhibits a model size that is 4,000 times smaller and demonstrates an inference speed that is ten times faster than that of its teacher counterpart.
- Through comprehensive evaluations using a large dataset of 10k student-written responses to science questions, our work not only validates the effectiveness of our KD method against state-of-the-art models like TinyBERT [7] and generic ANN models [3], but also highlights its superior performance.

## 2 Proposed Knowledge Distillation

KD is a technique to transfer knowledge from a trained large model (teacher) to a more compact and deployable model (student). We take inspiration from the prominent KD approach, introduced by [5], which involves using the class probabilities generated by the pre-trained large model as *soft labels* for training the smaller model, effectively transferring its predictive and generalization capabilities. Building on this concept, we develop a method for applying KD in the context of automated scoring systems, aiming to improve the process of evaluating educational content using AI.

Specifically, for each data point $\mathbf{x}_i$ in the training sample $\mathcal{D}$, the teacher model predicts the class probability $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})^T$, where $p_{ij}$ represents the predicted probability that the $i^{\text{th}}$ data point belongs to class $j$. The student model is trained using both the training sample $\mathcal{D}$ and the corresponding *soft*

*labels* $\boldsymbol{p} = \{\boldsymbol{p}_i\}_{i=1}^N$ produced by the teacher model. We represent the student model by a neural network $f(\cdot, \boldsymbol{\theta})$. The discrepancy between the student and teacher models is measured as

$$\begin{aligned} \tilde{\mathcal{L}}(f(\cdot, \boldsymbol{\theta}); \mathcal{D}, \boldsymbol{p}) &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}(\boldsymbol{p}_i, f(\mathbf{x}_i, \boldsymbol{\theta})), \\ &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \log \left( f_k(\mathbf{x}_i; \boldsymbol{\theta}) \right), \end{aligned} \tag{1}$$

which is the sample mean of the cross-entropy $\mathrm{CE}(\boldsymbol{p}_i, f(\mathbf{x}_i, \boldsymbol{\theta}))$ across $i$. To leverage the information from both the training data and the teacher model's predictions, KD aims to solve

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{KD}}^* &= \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \left\{ \mathcal{L}^{\mathrm{KD}}(f(\cdot, \boldsymbol{\theta}); \mathcal{D}, \boldsymbol{p}, \lambda) \right\} \\ &= \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \left\{ \mathcal{L}(f(\cdot, \boldsymbol{\theta}); \mathcal{D}) + \lambda \tilde{\mathcal{L}}(f(\cdot, \boldsymbol{\theta}); \mathcal{D}, \boldsymbol{p}) \right\}, \end{aligned} \tag{2}$$

where the minimized KD loss $\mathcal{L}^{\mathrm{KD}}(f(\cdot, \boldsymbol{\theta}); \mathcal{D}, \boldsymbol{p}, \lambda)$ is the linear combination of two loss terms in KD loss and (1). The first term of the KD loss equation measures the discrepancy between the predictions of the student model and the actual labels. The second term assesses the prediction discrepancy between the student and teacher models. In this context, $\lambda$ serves as a constant that balances the impact of these two aspects of the loss. Setting $\lambda = 0$ reduces the KD loss to the conventional empirical risk loss.

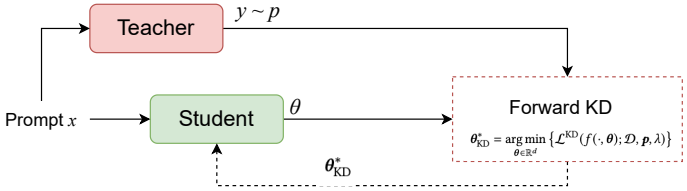


**Fig. 1.** The architecture of the proposed KD approach uses prediction probabilities as soft labels from the teacher model and forces the student model to achieve these prediction probabilities through the fitting loss function.

KD enables the student model to attain performance comparable to the teacher model while considerably reducing the computational resources required for training. The teacher model's predicted probability outputs $\boldsymbol{p}$ provide valuable insights into its data interpretation. By minimizing the discrepancy between the outputs of the student and teacher models, the student model can effectively adopt the knowledge and insights of the teacher model. Consequently, despite its simpler architecture and reduced computational resources, the student model can match the performance of the more complex teacher model.

In Fig. 1, we present the architecture of the proposed KD method. With a well-performing, fine-tuned, large teacher model and a new dataset, we run the teacher model on the dataset and extract knowledge to guide the training of a more compact student model. In this study, we extract the class probabilities

predicted by the teacher model as the knowledge to be transferred to the student model. sing both the knowledge from the teacher model and the data from the dataset, we train the student model based on optimization, as outlined in Equation (2).

## 3 Experimental Setup

Our study investigates whether a significantly smaller neural network can effectively mimic the capabilities of a fine-tuned LLM through the proposed KD strategy. Additionally, the study explores how this approach can enhance model performance. We apply our proposed methodology across diverse datasets to train a compact model to achieve this goal. This model is then compared with the SOTA TinyBERT [7] and a trained smaller ANN [3] to evaluate the performance in terms of accuracy and efficiency. TinyBERT stands out due to its specific design to compress the size and computational demands of BERT through a sophisticated distillation process, making it an ideal benchmark for assessing the efficiency and effectiveness of our distillation approach. On the other hand, ANN represents a broader category of neural networks that, while not specialized in natural language processing tasks to the same extent as TinyBERT, provides a contrasting baseline to evaluate the general applicability and performance of our distilled model in automatic scoring. The comparison against these models allows us to validate the superiority of our approach not only against a state-of-the-art distilled model like TinyBERT, which is directly relevant to our domain but also against more generic neural network architectures.

### 3.1 Data Collection and Preprocessing

The study utilized a meticulously categorized dataset of student-written responses to a science question and three mathematical assessment items, each falling under the multi-class category for automatic scoring. Each student response in datasets is graded by a human expert for automatic scoring, and human scores are used for validation. On average, each student's written textual response contains 15 words. The detailed composition of each assessment item's dataset is presented in Table 1.

| Dataset | Sample size | Classes | Teacher | Student |
|---|---|---|---|---|
| 7T | 6,684 | 10 | SciEdBERT (114M) | E-LSTM (0.03M) |
| Bathtub | 1,145 | 5 | BERT_base (110M) | E-LSTM (0.03M) |
| Falling Weights | 1,148 | 4 | BERT_base (110M) | E-LSTM (0.03M) |
| Gelatin | 1,142 | 5 | BERT_base (110M) | E-LSTM (0.03M) |

**Table 1.** Sample size and the teacher and student model used for each dataset. The number of parameters for each model is shown in parentheses.

| | Accuracy | | | |
|---|---|---|---|---|
| | Teacher | TinyBERT | ANN | KD |
| 7T | 0.891±0.016 | 0.752±0.003 | 0.716±0.002 | 0.757±0.001* |
| Bathtub | 0.938±0.014 | 0.833±0.019 | 0.831±0.021 | 0.852±0.012* |
| Falling Weights | 0.904±0.013 | 0.856±0.015 | 0.865±0.014 | 0.888±0.008* |
| Gelatin | 0.871±0.018 | 0.735±0.010 | 0.739±0.011 | 0.780±0.014* |
| | F-1 Score | | | |
| 7T | 0.842±0.017 | 0.749±0.009 | 0.706±0.001 | 0.751±0.005* |
| Bathtub | 0.914±0.021 | 0.832±0.069 | 0.830±0.024 | 0.851±0.011* |
| Falling Weights | 0.893±0.018 | 0.855±0.015 | 0.864±0.014 | 0.886±0.009* |
| Gelatin | 0.804±0.016 | 0.731±0.008 | 0.733±0.014 | 0.766±0.017* |

* KD has shown higher accuracy and F-1 score than TinyBERT and ANN, and is comparable to the Teacher model for each dataset.

**Table 2.** Accuracy and F-1 score performance comparison of teacher, TinyBERT [7], ANN [3], and KD model for benchmark datasets. The mean accuracies and standard deviations are displayed.

### 3.2 Dataset

The 7T dataset is a large dataset consisting of seven tasks from the SR1 dataset, including short constructed student responses and human-expert graded scores. Overall, the 7T dataset consists of 6,684 labeled student responses from [12], similar to the dataset used for SciEdBERT by Liu et al. [9]. We utilized three multi-class assessment tasks from the Mathematical Thinking in Science (MTS) project [2] responded to by high-school students: Bathtub, Falling Weights, and Gelatin containing 1,145, 1,148, and 1,142 student-written responses respectively. Each dataset contains a different number of classes (scores assigned by human experts) for student-written responses. More specific details about scoring and assessment items can be found in [8]. This comprehensive dataset facilitated a nuanced analysis of the capacity of compact scoring models for student-written responses, ensuring robust and broadly applicable study findings. We processed each dataset by excluding empty responses and ensuring text-formatted student responses and ranged labels.

### 3.3 Training Scheme

**Model Setup** This study uses SciEdBERT [9] with 114M parameters as a specialized Science Education BERT model, and the standard BERT base model [1] contains 110M parameters as the teacher model. These models have been shown to perform brilliantly in processing textual data. For performance comparison, we used TinyBERT [7] with 67M parameters and small ANN [3]. For the KD method, we constructed a compact neural network with an embedding layer with an output dimension of 32 and a bidirectional LSTM layer with 16 units (significantly fewer parameters than transformers), followed by a GlobalMax-Pooling1D layer. Further, it includes two dense layers, with the first having 16 neurons and 'relu' activation, and the final layer is equipped with a softmax activation for multi-class classification. Additionally, dropout layers are integrated for regularization, and the model is optimized with Adam.

**Evaluation and Validation** We partition each dataset into training, validation, and testing sets in a 7 : 1 : 2 ratio. The model optimization employs cross-entropy loss, and to prevent overfitting, an early stopping callback that monitors the validation loss is utilized. We present the prediction accuracy on the test set to assess the model's performance.

The summary of the dataset and the teacher and student (KD) models used for each dataset is detailed in Table 1. We provide the number of parameters for each model in parentheses. The student model is much smaller than the teacher model.

### 3.4 Results

The comparative analysis of model accuracy across four datasets is presented in Table 2. Results reveal the efficacy of KD in enhancing the performance of a student model as compared to the SOTA TinyBERT [7] and ANN [3] for text classification, in terms of both accuracy and F-1 score. Furthermore, it also provides close accuracy and F-1 score as the complex teacher model. The Falling Weights dataset serves as a typical example, with KD providing performance comparable to the teacher model, suggesting that even models with much smaller sizes can achieve similar performance to the large teacher model. We observed that KD outperforms TinyBERT and ANN in accuracy by 2.5% and 3.2%, respectively, and in F-1 score by 2.2% and 3.0%, respectively. This observation highlights the superiority of KD over SOTA model distillation approaches. Considering both accuracy, F-1 score (shown in Table 2), and model size (shown in Table 1), results highlight the practicality and applicability of the KD approach for automatic scoring on resource constrained-devices.

Despite the success of KD, it is essential to recognize that the student models, although improved, usually do not reach the performance benchmarks set by the teacher models. This is notably apparent in the 7T dataset; the integration of KD leads to better performance compared to the ANN and TinyBERT but still does not match the teacher models' accuracy. Such a discrepancy can be attributed to the inherent limitations of the student models, which possess simpler architectures and are trained on smaller datasets with far fewer training parameters.

The results demonstrate that the KD strategy is a powerful tool in model training, beneficial for applications such as automatic scoring. By effectively condensing the knowledge of a large, pre-trained model into a more compact one, KD not only improves performance but also facilitates the deployment of such models in resource-constrained environments.

### 3.5 Sensitivity Analysis

We investigated the impact of the hyperparameter $\lambda$ in Eq. (2) on scoring accuracy to learn more about the resilience of the KD approach. We assessed the KD approach using a step size of 0.02 and a range of $\lambda$ values from 0.08 to 0.02 for the Bathtub dataset. Although there were little variations in the accuracy of the

KD technique with the modification of $\lambda$, consistently outperformed the baseline models. These findings suggest that the KD approach is comparatively resistant to the selection of $\lambda = 0.2$, demonstrating consistent performance across a range of hyperparameter values.

## 4 Discussion

The results of this study highlight the revolutionary possibilities of KD in educational technology, especially in light of the limitations of standard school computing resources. The use of KD in education represents a substantial breakthrough, particularly in automated grading systems. Nevertheless, like any emerging technology, it is important to recognize its limitations as well as its potential for development in the future. In the traditional education system, automatic evaluation is yet a point of discussion [10]. Therefore, our proposed solution is a supplementary tool designed to support and not replace traditional assessment methods established in the education system. Further studies and educational policy adaptations are necessary to fully integrate such technologies into formal school environments.

The most noteworthy application of KD in education is the creation of accurate and efficient automatic scoring systems. A major challenge in many educational contexts is that traditional scoring systems can demand extensive processing resources to function successfully on school-setting devices such as entry-level laptops and tablets. This problem is addressed by KD, which enables the creation of "student models" with significantly lower processing requirements while preserving much of the accuracy and efficiency of larger "teacher models.

Furthermore, KD models are ideally suited for integration into tablet- and smartphone-based learning apps due to their smaller size and reduced processing requirements. The capacity to run complex AI models on these devices, which are increasingly prevalent in educational contexts, creates new opportunities for interactive and adaptable learning experiences.

## 5 Conclusion

This study effectively illustrates how KD can be used to optimize LLMs for usage in educational technology, especially on low-processor devices. We maintain great accuracy with a much smaller model size (0.03M parameters) and processing requirements by condensing the knowledge of LLMs into smaller neural networks. The distilled models perform better than SOTA TinyBERT and ANN models on various datasets, demonstrating the efficacy of this approach even though their parameter sizes are up to 100 times less than teacher models. This work has important applications since it provides a method to incorporate cutting-edge AI tools into conventional school environments, which frequently have hardware constraints. The learning process and accessibility of personalized education technology can be significantly improved by the capacity to implement effective and precise AI models for uses such as autonomous scoring. Essentially,

this work establishes the foundation for future developments in the field and validates the viability of KD in educational contexts, underscoring the significance of ongoing research and innovation in AI for education. In the future, we will work on processing soft-labels and prompt processing to avoid amplification of faults of teacher models by employing more sophisticated techniques.

## Acknowledgment

This work was funded by the National Science Foundation(NSF) (Award Nos. 2101104, 2138854) and partially supported by the NSF under grants DMS-1903226, DMS-1925066, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809, and by the U.S. National Institutes of Health under grant R01GM152814.

## References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. ETS-MTS, T.: Learning progression-based and ngss-aligned formative assessment for using mathematical thinking in science. http://ets-cls.org/mts/index.php/assessment/ (November, 2023)
3. Ghiassi, M., Olschimke, M., Moon, B., Arnaudo, P.: Automated text classification using a dynamic artificial neural network model. Expert Systems with Applications **39**(12), 10967–10976 (2012)
4. González-Calatayud, V., Prendes-Espinosa, P., Roig-Vila, R.: Artificial intelligence for student assessment: A systematic review. Applied Sciences **11**(12), 5467 (2021)
5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
6. Holmes, W., Tuomi, I.: State of the art and practice in ai in education. European Journal of Education **57**(4), 542–570 (2022)
7. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
8. Latif, E., Zhai, X.: Fine-tuning chatgpt for automatic scoring. Computers and Education: Artificial Intelligence p. 100210 (2024)
9. Liu, Z., He, X., Liu, L., Liu, T., Zhai, X.: Context matters: A strategy to pre-train language model for science education. arXiv preprint arXiv:2301.12031 (2023)
10. Selwyn, N.: Should robots replace teachers?: AI and the future of education. John Wiley & Sons (2019)
11. Zhai, X.: Chatgpt user experience: Implications for education. Available at SSRN 4312418 (2022)
12. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. Journal of Research in Science Teaching **59**(10), 1765–1794 (2022)
13. Zhai, X., Yin, Y., Pellegrino, J.W., Haudek, K.C., Shi, L.: Applying machine learning in science assessment: a systematic review. Studies in Science Education **56**(1), 111–151 (2020)
14. Zhai, X., Chu, X., Chai, C.S., Jong, M.S.Y., Istenic, A., Spector, M., Liu, J.B., Yuan, J., Li, Y.: A review of artificial intelligence (ai) in education from 2010 to 2020. Complexity **2021**, 1–18 (2021)