Efficient Multi-task Inferencing: Model Merging with Gromov-Wasserstein Feature Alignment

Luyang Fang[1,3], Ehsan Latif[1,2], Haoran Lu[3], Yifan Zhou[1,4], Ping Ma[3], and Xiaoming Zhai[1,2(✉)]

[1] AI4STEM Education Center, Athens, GA, USA
xiaoming.zhai@uga.edu
[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
[4] School of Computing, University of Georgia, Athens, GA, USA

Abstract. Automatic scoring of student responses enhances efficiency in education, but deploying a separate neural network for each task increases storage demands, maintenance efforts, and redundant computations. To address these challenges, this paper introduces the Gromov-Wasserstein Scoring Model Merging (GW-SMM) method, which merges models based on feature distribution similarities measured via the Gromov-Wasserstein distance. Our approach begins by extracting features from student responses using individual models, capturing both item-specific context and unique learned representations. The Gromov-Wasserstein distance then quantifies the similarity between these feature distributions, identifying the most compatible models for merging. Models exhibiting the smallest pairwise distances, typically in pairs or trios, are merged by combining only the shared layers preceding the classification head. This strategy results in a unified feature extractor while preserving separate classification heads for item-specific scoring. We validated our approach against human expert knowledge and a GPT-o1-based merging method. GW-SMM consistently outperformed both, achieving higher micro F1 score, macro F1 score, exact match accuracy, and per-label accuracy. The improvements in micro F1 and per-label accuracy were statistically significant compared to GPT-o1-based merging ($p = 0.04$, $p = 0.01$). Additionally, GW-SMM reduced storage requirements by two-thirds without compromising much accuracy, demonstrating its computational efficiency alongside reliable scoring performance.

Keywords: Multi-task learning · Efficient inference · Automatic Scoring · Deep Learning · Optimal Transport

L. Fang and E. Latif—Co-first authors.

A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 192–200, 2025.
https://doi.org/10.1007/978-3-031-99267-4_24

1 Introduction

Automated scoring systems have become indispensable in modern educational assessment, enabling efficient evaluation of students' open-ended responses, particularly in science and STEM domains [15,17]. As curriculum frameworks such as the Next Generation Science Standards (NGSS) promote complex performance tasks to assess multidimensional understanding [9], the demand for reliable and scalable automated scoring continues to rise. Deep learning models like BERT [3], offer robust performance in such tasks. However, deploying a separate fine-tuned model for each assessment item is often impractical due to high memory and inference costs—especially in resource-constrained environments such as browser-based student assessments or serverless educational platforms [29].

In large-scale assessments, models are frequently fine-tuned separately per task to align with domain-specific rubrics and expert scoring criteria [22]. Over time, educational organizations may accumulate numerous task-specific models. Unlike multitask learning setups where tasks are trained jointly from scratch, re-training or co-training becomes infeasible due to data silos, annotation inconsistencies, and vendor diversity [14]. While parameter-efficient tuning methods like LoRA [11], adapters [4], or knowledge distillation [5,8] offer alternatives, they require retraining access and architectural constraints that are not always practical in real-world deployments.

The problem we address is thus real and pressing: *How can we consolidate multiple pre-trained scoring models, each trained on different tasks, into a smaller set of merged models without causing significant performance degradation?* This scenario frequently arises in automated essay scoring (AES) and short-answer grading (ASAG), particularly in cross-prompt trait scoring settings [12,28]. Prior works have explored joint training across prompts, transfer learning across similar tasks [16], and meta-learning for prompt generalization [7], but the question of post hoc model merging remains largely underexplored.

To bridge this gap, we introduce *Gromov-Wasserstein Scoring Model Merging (GW-SMM)*, a new algorithm that merges pre-trained models by aligning their learned response representations. In contrast to prior methods that rely on parameter similarity or prompt metadata, GW-SMM uses the Gromov-Wasserstein (GW) distance, a structure preserving optimal transport metric [19,24], to measure alignment between model-specific student response features. This strategy ensures that only models with structurally compatible internal representations are merged. It minimizes conflicts and maximizes knowledge transfer.

The proposed method supports resource-efficient AI for education and other cost-sensitive domains, contributing to advancements in scalable, efficient model deployment. Our key contributions are summarized below:

– We propose a novel model merging approach using GW distance to align deep semantic representations of student responses, enabling scalable deployment of scoring models without retraining.

– We empirically validate GW-SMM on NGSS-aligned science assessments across nine tasks, demonstrating significant storage reduction with minimal performance trade-offs, outperforming both human knowledge-based and GPT-derived merging strategies.
– We also open-source the code on Github[1] repository for reproducibility.

2 Gromov-Wasserstein Scoring Model Merging

We propose Gromov-Wasserstein Scoring Model Merging (GW-SMM) to merge multiple task-specific scoring models into a reduced set while preserving performance. Given T fine-tuned models trained on distinct NGSS-aligned science assessments, GW-SMM merges models by aligning their learned semantic features of student responses. The workflow is shown in Fig. 1.

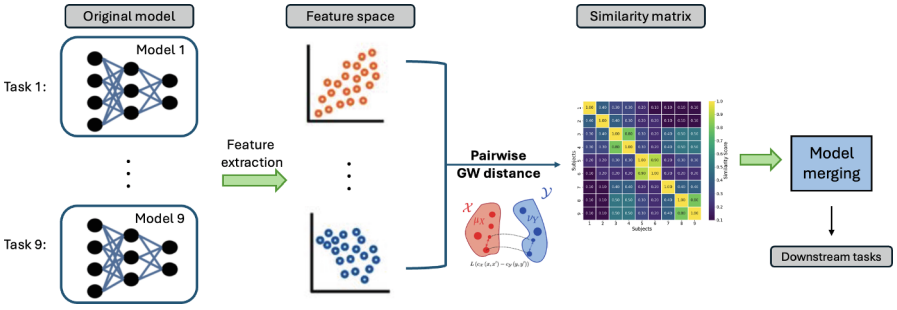

Fig. 1. Overview of proposed GW-SMM, where models for tasks are merged based on the inter-feature GW distance.

Task-Specific Representation Extraction. For each fine-tuned automated scoring model, we extract high-dimensional semantic features of student responses by passing task-specific data through the corresponding model and collecting the activations from the final hidden layer (pre-classification). In transformer architectures (e.g., BERT), this layer captures task-specific semantic patterns critical for scoring decisions, as shown in prior work on model similarity analysis [3,20].

Structural Alignment via Gromov-Wasserstein Distance. To identify which models should be merged together, we assess similarity based on their extracted features, which reflect both student response information and underlying model structures. We compute pairwise distances between their feature spaces using the GW distance. The GW distance captures the geometric alignment of features, allowing comparison even when models encode responses in heterogeneous metric spaces, such as those with different dimensionalities or context-dependent semantics.

[1] https://github.com/AI4STEM-Education-Center/MoE.git.

Formally, let $\mathbf{X} \in \mathbb{R}^{n_i \times d_i}$ and $\mathbf{Y} \in \mathbb{R}^{n_j \times d_j}$ denote the extracted feature for model i and j, where n_i, n_j and d_i, d_j may differ across models. Note that we work on a special case that each fine-tuned model has the same model architecture, i.e., $d_i = d_j$, but the framework is generally feasible for settings with different d_i and d_j. Let $C_i \in \mathbb{R}^{n_i \times n_i}$ and $C_j \in \mathbb{R}^{n_j \times n_j}$ denote intra-model cost matrices where $C_i[k,l] = distance_i(\mathbf{x}_k, \mathbf{x}_l)$ and $C_j[m,n] = distance_j(\mathbf{y}_m, \mathbf{y}_n)$. To align the feature structures between two models, we introduce a coupling matrix $\pi \in \mathbb{R}_+^{n_i \times n_j}$, where $\pi_{k,m}$ indicates the matching strength between the k-th point in $\mathbf{X}$ and the m-th point in $\mathbf{Y}$. The GW alignment is computed through the coupling matrix $\pi \in \mathbb{R}_+^{n_i \times n_j}$ solving:

$$GW(C_i, C_j)^2 = \min_{\pi \in \Pi(\mathbf{p},\mathbf{q})} \sum_{k,l=1}^{n_i} \sum_{m,n=1}^{n_j} \pi_{k,m}\pi_{l,n} L(C_i[k,l], C_j[m,n]), \tag{1}$$

where L is a distance measure with common choices and $L(a,b) = |a-b|^p$, $p \geq 1$, and $\Pi(\mathbf{p},\mathbf{q}) = \left\{\pi \mid \pi_{k,m} \geq 0; \forall k, \sum_{m=1}^{n_j} \pi_{k,m} = p_k; \forall m, \sum_{k=1}^{n_i} \pi_{k,m} = q_m\right\}$ enforces marginal constraints with $\mathbf{p} \in \mathbb{R}^{n_i}, \mathbf{q} \in \mathbb{R}^{n_j}$ as weights for points in $\mathbf{X}$ and $\mathbf{Y}$ (typically uniform: $p_k = 1/n_i, q_m = 1/n_j$).

Direct optimization of GW is challenging due to its non-convex and non-smooth properties. Thus, we use entropy regularization [23], which renders the problem convex while maintaining geometric fidelity between feature spaces, with the optimal coupling computed via Sinkhorn iterations [2,25]. For theoretical guarantees, computational considerations, and extensions of the discrete GW formalism, we direct readers to [21,24].

Determination of the Merging Plan. Given the distance matrix and the target number T' of final merged models, we determine the optimal merging plan by maximizing the total similarity, $L(\mathcal{M}) = \sum_{(i,j)\in\mathcal{M}} S_{i,j}$, where $\mathcal{M}$ denotes the set of merged pairs or groups, and $S_{i,j}$ is the normalized similarity between models i and j, defined as $S_{i,j} = 1 - \frac{d_{i,j} - d_{\min}}{d_{\max} - d_{\min}}$, where $d_{i,j}$ is the GW distance between models i and j, and $d_{\min}$, $d_{\max}$ are the minimum and maximum distances. We search for the optimal merging plan $\mathcal{M}^*$ that minimizes $L(\mathcal{M})$ over all possible merge plans that will result in T' final merged models.

Model Merging. With the designed merging plan, we adopt `TIES-MERGING` [28] to merge task-specific models into a unified model while preserving performance. Suppose we have T models that will be merged together. For model $t \in \{1, \cdots, T\}$, we compute the parameter updates $\tau_t = \theta_t - \theta_0$, where θ_t denotes the parameters for the task-specific model t while θ_0 denotes the parameters for the base model. With an alignment and pruning process to remove redundant parameters and conflicting parameters across different models via magnitude-based sparsification, we have the updated updates γ_t, and the final merged parameters are $\theta_{\text{merged}} = \theta_0 + \sum_{t=1}^{T} \lambda_t \gamma_t$, where the coefficients λ_t are determined by the alignment and pruning process. By combining representation alignment with targeted pruning, `TIES-MERGING` effectively leverages shared knowledge across tasks while maintaining task-specific distinctions.

3 Dataset Details

This study utilizes expert-scored responses from approximately 1,200 middle school students across the U.S., completing nine NGSS-aligned, open-ended science assessment tasks from the PASTA project [9,22]. These multi-label tasks, adapted from the Next Generation Science Assessment (NGSA) initiative, target physical science concepts under "Matter and Its Characteristics" domain and require students to integrate disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs) [1]. Responses were anonymized and collected in diverse educational contexts, reflecting variability in instructional methods, digital access, and linguistic backgrounds. For example, in one task, students analyzed density and flammability data to identify unknown gases, applying SEP, CCC, and DCI reasoning. All responses were scored using a structured multi-dimensional rubric co-designed with educators, enabling nuanced evaluation of students' scientific thinking [10]. Detailed dataset statistics and scoring rubrics are available on the project repository[1].

4 Experimentation and Results

For each task, we have a fine-tuned BERT-base model [3] for multi-label classification of student responses, tokenized via WordPiece [27]. Each response was mapped to a fixed-length sequence with [CLS]/[SEP] tokens and encoded by BERT. Training used AdamW [18] (learning rate: $2e^{-5}$, batch size: 32) with binary cross-entropy loss per label and dropout regularization [26]. Early stopping was applied based on validation loss. The training process was conducted for varying numbers of epochs (ranging from 10 to 20) on an NVIDIA GPU. For merging, we removed the classification head and extracted [CLS] token embeddings (dimension $d = 768$) from the final layer, forming a matrix (n_i, d) for each task, where n_i is the total number of student responses. Pairwise Euclidean distances between features were computed to inform model merging decisions.

We evaluate each method using four multi-label metrics: Micro F1 (overall performance weighted by label frequency), Macro F1 (equal weight across labels), Exact Match (all labels must match exactly), and Per-label Accuracy (accuracy per label across instances). Together, they capture both overall and label-specific performance. Performance is averaged across nine tasks, using merged or original models as applicable. We compare GW-SMM with two baselines: **(1) Human Knowledge:** Merging plans based on task similarities identified by domain experts through content and rubric analysis - reflecting common practices in educational assessment; **(2) GPT-o1:** Merging plans are generated by prompting GPT-o1 to compare task descriptions, inspired by LLM-based semantic similarity evaluation methods. As shown in Gatto et al. [6], such methods achieve strong performance in similarity tasks, making them reliable benchmarks.

We note that GW-SMM is a data-driven approach that leverages all student responses to infer task relationships, whereas Human Knowledge and GPT-o1 rely solely on task descriptions. This distinction stems from practical and computational constraints: human experts cannot feasibly analyze large volumes of

student responses manually, and GPT-o1 faces similar limitations due to data privacy concerns and the high cost of processing large-scale data [13].

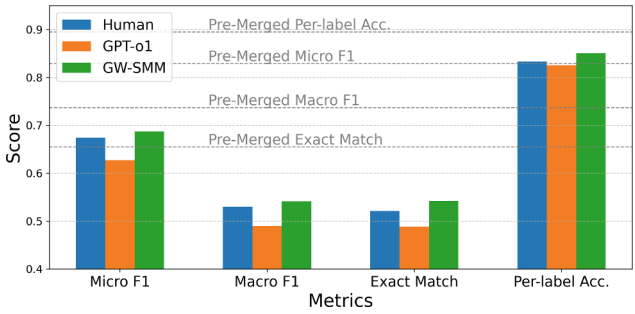

Fig. 2. Performance results before and after merging using three methods (GW-SMM (Ours), human knowledge, and GPT-o1).

Table 1. Statistical comparison of methods across metrics. Each value shows the t-statistic and p-value for pairwise comparisons among GW-SMM, Human Knowledge, and GPT-o1. Statistically significant results ($p < 0.05$) are bolded.

Metrics	GW-SMM vs Human		GW-SMM vs GPT-o1		Human vs GPT-o1	
	T-Stat	P-Value	T-Stat	P-Value	T-Stat	P-Value
Micro F1 Score	0.869983	0.391454	**2.106563**	**0.043925**	1.190602	0.243465
Macro F1 Score	-0.187336	0.852702	1.273012	0.213125	1.604841	0.119365
Exact Match Accuracy	1.363881	0.183098	1.173100	0.250302	0.062403	0.950670
Per-label Accuracy	1.577693	0.125483	**2.637638**	**0.013281**	0.769699	0.447703

In this study, we merged original models into $T' = 3$ merged models, reducing storage costs by 3x. As shown in Fig. 2, GW-SMM outperforms other methods, effectively aggregating models while retaining accuracy. Despite degradation versus the Pre-Merge baseline, GW-SMM achieves the optimal balance of efficiency and accuracy. Merging strategies and full results across metrics and tasks are available in our repository[5].

We also conducted paired t-tests to compare the effectiveness of each method (Table 1). The results reveal that GW-SMM and GPT-o1 exhibit statistically significant differences on two metrics: Micro F1 Score ($t = 2.107$, $p = 0.044$) and Per-label Accuracy ($t = 2.638$, $p = 0.013$). These results indicate that GW-SMM outperforms GPT-o1 in these aspects. In contrast, no statistically significant differences ($p > 0.05$) were observed between GW-SMM and Human Knowledge or between Human Knowledge and GPT-o1 across any metric. For instance,

comparisons between GW-SMM and Human Knowledge yielded p-values ranging from 0.126 to 0.853, while Human Knowledge versus GPT-o1 resulted in p-values between 0.120 and 0.951. These findings suggest that GW-SMM effectively leverages model merging to enable robust and efficient scoring, performing comparably to, or even exceeding, Human Knowledge in some cases.

5 Conclusion

We propose GW-SMM, a novel model merging method for the scalable deployment of commonly featured automatic scoring models. The method uses the Gromov-Wasserstein distance to align task-specific feature spaces by measuring structural divergence in student response features. GW-SMM identifies compatible models for merging while minimizing conflicts. Evaluated on nine NGSS-aligned science tasks, GW-SMM outperformed both human expert-guided and GPT-o1-based merging strategies, achieving statistically significant improvements in micro F1 score (0.6872 vs. 0.6271, $p = 0.04$) and per-label accuracy (0.8507 vs. 0.8255, $p = 0.01$). The method reduced storage costs by up to 3× while maintaining competitive performance relative to standalone models. Its data-driven approach, which leverages response patterns rather than task descriptions, ensures adaptability to diverse student responses. Future work will focus on advanced fusion techniques to further narrow the performance gap with pre-merged models. GW-SMM advances scalable, efficient deployment of AI in education, balancing accuracy and resource constraints.

Acknowledgments. This work was partially supported by the Institute of Education Sciences (IES) [R305C240010], the U.S. National Science Foundation (NSF) [DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809], and the National Institutes of Health [NIH R01GM152814]. Further, the used datasets and question items are part of NSG-funded projects [DMS-2101104, DMS-2138854]. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, or IES.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Council, N.R., et al.: Next generation science standards: for states, by states (2013)
2. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, vol. **26** (2013)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805 (2018)
4. Ding, N., et al.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat. Mach. Intell. **5**(3), 220–235 (2023)
5. Fang, L., Chen, Y., Zhong, W., Ma, P.: Bayesian knowledge distillation: a Bayesian perspective of distillation with uncertainty quantification. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)

6. Gatto, J., Sharif, O., Seegmiller, P., Bohlman, P., Preum, S.M.: Text encoders lack knowledge: Leveraging generative LLMs for domain-specific semantic textual similarity. arXiv preprint arXiv:2309.06541 (2023)
7. Geden, M., Emerson, A., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in educational games with multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 654–661 (2020)
8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vis. **129**(6), 1789–1819 (2021)
9. Harris, C.J., Krajcik, J.S., Pellegrino, J.W.: Creating and using instructionally supportive assessments in NGSS classrooms. NSTA Press (2024)
10. He, P., Shin, N., Zhai, X., Krajcik, J.: Guiding teacher use of artificial intelligence-based knowledge-in-use assessment to improve instructional decisions: a conceptual framework. In: Zhai, X., Krajcik, J. (eds.) Uses of Artificial Intelligence in STEM Education. Oxford University Press (2024)
11. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
12. Katuka, G.A., Gain, A., Yu, Y.Y.: Investigating automatic scoring and feedback using large language models. arXiv preprint arXiv:2405.00602 (2024)
13. Koedinger, K.R., D'Mello, S., McLaughlin, E.A., Pardos, Z.A., Rosé, C.P.: Data mining and education. Wiley Interdisc. Rev. Cogn. Sci. **6**(4), 333–353 (2015)
14. Koprinska, I., et al.: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings. Part II, Springer Nature (2023). https://doi.org/10.1007/978-3-031-74640-6
15. Latif, E., Zhai, X.: Fine-tuning ChatGPT for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100210 (2024)
16. Latif, E., Zhai, X.: Efficient multi-task inferencing with a shared backbone and lightweight task-specific adapters for automatic scoring. In: AAAI Workshop on iRAISE: Innovation and Responsibility in AI-Supported Education, pp. 1–12. PMLR (2025)
17. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100213 (2024)
18. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Mémoli, F.: The Gromov-Wasserstein distance: a brief overview. Axioms **3**(3), 335–341 (2014)
20. Merchant, A., Rahimtoroghi, E., Pavlick, E., Tenney, I.: What happens to BERT embeddings during fine-tuning? arXiv preprint arXiv:2004.14448 (2020)
21. Panaretos, V.M., Zemel, Y.: Statistical aspects of Wasserstein distances. Ann. Rev. Stat. Appl. **6**(1), 405–431 (2019)
22. Pasta, P.T.: Supporting instructional decision making: potential of an automatically scored three-dimensional assessment system. https://ai4stem.org/pasta/ (2023)
23. Peyré, G., Cuturi, M., Solomon, J.: Gromov-Wasserstein averaging of kernel and distance matrices. In: International Conference on Machine Learning, pp. 2664–2672. PMLR (2016)
24. Peyré, G., Cuturi, M., et al.: Computational optimal transport: with applications to data science. Found. Trends® Mach. Learn. **11**(5-6), 355–607 (2019)
25. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pac. J. Math. **21**(2), 343–348 (1967)

26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learning Rres. **15**(1), 1929–1958 (2014)
27. Wu, Y.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
28. Yadav, P., Team: Ties-merging: a pruning-based approach to alleviate conflicts in model merging. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 4567–4578 (2023)
29. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)