One Model to Score Them All: Unified Scoring of Learning Strategies with LLMs

Andreea Dutulescu
National University of Science and Technology Politehnica Bucharest
Academy of Romanian Scientists, Bucharest
andreea.dutulescu@upb.ro

Stefan Ruseti
National University of Science and Technology Politehnica Bucharest
Academy of Romanian Scientists, Bucharest
stefan.ruseti@upb.ro

Mihai Dascalu
National University of Science and Technology Politehnica Bucharest
Academy of Romanian Scientists, Bucharest
mihai.dascalu@upb.ro

Danielle McNamara
Arizona State University
dsmcnama@asu.edu

ABSTRACT

The assessment of student responses to learning-strategy prompts, such as self-explanation, summarization, and paraphrasing, is essential for evaluating cognitive engagement and comprehension. However, manual scoring is resource-intensive, limiting its scalability in educational settings. This study investigates the use of Large Language Models for automating the evaluation of student responses based on expert-defined rubrics. We fine-tune open-source LLMs on annotated datasets to predict expert ratings across multiple scoring rubrics, ensuring consistency and efficiency in assessment. Our findings indicate that multi-task fine-tuning, which involves training a single model across multiple scoring tasks, consistently outperforms single-task training by enhancing generalization and mitigating overfitting. This advantage is particularly noticeable in recent architectures, where multi-task training enables robust performance across diverse evaluation criteria. Notably, our Llama 3.2 3B model achieved high performance, outperforming a 20x larger zero-shot model while maintaining feasibility for deployment on consumer-grade hardware, emphasizing the potential for scalable AI-driven assessment solutions. This research contributes to open education by fine-tuning open-source models and publicly releasing trained models, training scripts, and evaluation frameworks. The proposed approach supports automated, reproducible, and scalable assessment of learning strategies, facilitating timely feedback for students and reducing the burden on educators. `https://github.com/upb-nlp/EDM-LLM-Scoring`

Keywords

learning strategies, scoring, Large Language Models

Andreea Dutulescu, Stefan Ruseti, Mihai Dascalu, and Danielle Mcnamara. One Model to Score Them All: Unified Scoring of Learning Strategies with LLMs. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 496–502. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870258

1. INTRODUCTION

Effective learning strategies (e.g., self-explanations, think-alouds, summaries, and paraphrases) play a critical role in shaping a student's ability to comprehend and retain information. These strategies encourage students to actively engage with the learning material by reflecting, articulating, and reorganizing knowledge, fostering deeper understanding and cognitive development. Students refine their skills iteratively by leveraging active learning techniques. As a result, evaluating the quality of students' use of these strategies is essential in educational research and pedagogy. However, manually assessing student responses to various learning strategies is time-consuming and prone to subjective bias in human judgment, making it difficult to scale such evaluations in classroom practice.

In recent years, Large Language Models (LLMs) have opened new opportunities to automate tasks that require a contextualized semantic analysis of student productions. Leveraging LLMs offers a solution to automatically score responses to learning strategies while maintaining consistency and efficiency. This is particularly beneficial in educational settings where feedback on learning strategies, such as self-explanation and summarization, can enhance student performance and foster metacognitive skills. Automating the scoring process enables tutors to provide timely feedback at scale, support formative assessment practices, and free up resources for more personalized instruction. Moreover, students can engage in multiple learning strategies, refine their approach, and iteratively improve by receiving immediate feedback on their ability to use these learning strategies.

In this work, we focus on experimenting with LLMs to evaluate student responses using instructions designed to elicit various learning strategies. These instructions may ask students to summarize a text, paraphrase key ideas, articulate their reasoning process via think-alouds, or engage in self-explanation to clarify their understanding. Expert educators often assess these responses based on well-defined rubrics, which may consider factors such as cohesion, wording quality, paraphrase quality, and use of domain-specific terminology.

We fine-tune LLMs on datasets containing student responses along with their ratings given by experts. Besides predicting these ratings, we aim to develop models that can replicate human judgments and generalize across a variety of learning strategies. We release open-source models that can reliably score student responses across multiple rubrics, providing an automated, scalable solution for evaluating learning strategies in educational settings.

Our results show that even reduced-size LLMs, which require significantly less computational power than their larger counterparts, perform remarkably well for scoring. This finding is particularly important because these models can be run on consumer-grade hardware, enabling widespread use in diverse educational settings without costly infrastructure. Our approach addresses several key challenges in the field as we explore the capability of LLMs to capture the nuanced differences in student responses. We contribute to the ongoing effort in educational research to enhance the feedback loop in learning environments, making it possible for students to receive targeted insights on their use of learning strategies.

In summary, this paper's contributions can be outlined as follows:

- We fine-tune Large Language Models to automate the scoring of student responses to various learning-strategy instructions. Our models achieve high performance on a wide range of rubrics, arguing for the potential to handle diverse educational tasks;
- We show that multi-task fine-tuning outperforms single-task fine-tuning across the majority of scoring rubrics;
- We contribute to open education by using and fine-tuning open-source models and providing open access to the models, training scripts, and evaluation frameworks, thus facilitating reproducibility and further research.

2. RELATED WORK

Early efforts in automating the evaluation of student responses relied on handcrafted linguistic features to predict the designed rubrics. Starting from the early studies of Page [17] with limited capabilities to capture the full complexity of human language, researchers considered handcrafted features to capture linguistic and content-based elements of student writing. These systems typically relied on surface-level features, including word counts, sentence length, and keyword matching, to assess the quality of student responses. Techniques such as part-of-speech tagging, syntactic parsing, and discourse analysis were employed to provide a richer understanding of student language. Ruseti et al. [19] introduced a hybrid model combining recurrent neural networks (RNNs) and textual complexity indices to score summaries automatically. The proposed system achieved adequate accuracy in a classification task that assessed how well the main ideas of the original text were captured in the summary. However, these systems could not deeply understand the meaning of student responses, making them somewhat limited in generalizing across diverse tasks and contexts.

Unlike earlier models that relied on handcrafted features or shallow language representations, Transformers enabled a much deeper understanding of text by capturing contextual information across entire sentences and passages. Models fine-tuned on specific educational datasets generalized more effectively across tasks, outperforming traditional feature-based methods. Botarleanu et al. [1, 2] investigated how well summaries captured the main idea of a reference text. The trained models incorporated domain-independent textual complexity indices, along with Transformers for semantic contextualization. This approach achieved low errors, outperforming bag-of-word representations or models considering only linguistic features. Moreover, their results suggested that combining linguistic and semantic indices leads to accurate and robust summary evaluations. Nicula et al. [16] explored paraphrase quality assessment. Focusing on four dimensions (i.e., lexical, syntactical, semantic, and overall quality) the authors approach the task by combining handcrafted features, siamese neural networks, and pre-trained BERT models with transfer learning from a larger paraphrase corpus. The results suggest that the models with transfer learning from general paraphrase datasets achieve enhanced performance.

The latest advancement in automated scoring of student responses comes from the use of Large Language Models, which have shown an unprecedented capability to comprehend and follow instructions across a wide range of tasks. For example, Nicula et al. [15] explored using LLMs, specifically FLAN-T5, to assess students' self-explanations automatically. The study evaluated LLMs in two scenarios: 0-shot and fine-tuned. In the 0-shot scenario, GPT3.5-turbo excelled in overall quality and improved with more examples in the prompt. In the fine-tuning scenario, FLAN-T5 models were fine-tuned using LoRA, resulting in significant performance boosts. Performance scaled with model size and the number of prompt examples. Song et al. [20] addressed the limitations of previous methods that required extensive fine-tuning on large datasets for essay-scoring tasks. They explored few-shot prompting and prompt tuning on a small dataset of student essays. The findings reveal that open-source LLMs, particularly those with 10B parameters, achieve performance comparable to some fine-tuned deep-learning baselines and can be further enhanced with optimized prompts.

A novel approach that involves fine-tuning small encoder models for multi-rubric essay scoring is studied by Wang et al. [21]. They propose a Mixture-of-Experts model to improve multi-trait scoring effectiveness. The approach represents essays using a pre-trained encoder-based model, where a gating mechanism directs token representations to specialized trait-specific experts. Each expert, implemented as a fully connected layer, learns distinct trait representations, with final scores predicted using a sigmoid activation function. To improve representation diversity and capture inter-trait relationships, the method introduces three regularization strategies: scoring diversity regularization, which ensures experts focus on unique traits; trait representation correlation regularization, leveraging contrastive learning to refine trait correlations; and trait correlation loss, aligning predicted and actual scores.

A model introduced for zero-shot evaluation is Prometheus 2 [8]. The authors fine-tuned an evaluator designed specifically to score model-generated text rather than student responses. Prometheus 2 unifies two model-based evaluation paradigms (i.e., direct assessment and pairwise ranking) by merging the weights of two LLMs separately trained on these formats. The key insight is that this weight-merging approach outperforms both jointly trained and single-format evaluator LLMs. The model is trained using a feedback dataset for direct assessment and a preference dataset, with evaluation criteria for helpfulness, harmlessness, and more. Using Mistral-7B [6] and Mixtral-8x7B [7] as base models, Prometheus 2 achieves the highest correlation with human evaluators and proprietary LLM-based judges across multiple benchmarks.

3. METHOD

3.1 Datasets and Scoring Tasks

We considered four scoring tasks corresponding to different learning strategies in our experiments: self-explanation, think-aloud, summarization, and paraphrasing. For each task, we rely on expert-annotated datasets, which provide high-quality labels to assess student performance based on predefined rubrics. These tasks were chosen for their relevance in evaluating key learning strategies and their impact on comprehension. This section outlines the datasets used in our study, the annotations, and the overall scoring methodology. For more detailed information on the specific scoring rubrics and task instructions, please refer to the code repository.

Self-explanation is a cognitive process that involves actively explaining the meaning of the text to oneself while reading. Research has shown that effective self-explanation can significantly improve reading comprehension. McNamara [11, 12, 13] showed that providing students with reading strategy instruction and self-explanation practice can enhance their understanding of complex texts, especially for those with lower domain knowledge. Paraphrasing and bridging are key strategies used in self-explanations to improve comprehension.

Think-aloud is a strategy where readers verbalize their thoughts and understanding while reading a text. By voicing their thoughts, readers can assess whether they grasp the main ideas. Think-aloud can help readers connect new information to their existing knowledge while encouraging them to participate actively in the reading process.

For self-explanation and think-aloud tasks, we use the dataset constructed by McNamara et al. [13], where experts rated student responses when prompted to self-explain certain sentences from a text. The responses were graded on several rubrics for both tasks: paraphrase presence, lexical change, syntactic change, misconception presence, monitoring, bridge presence, bridge contribution, elaboration presence, life event recollection, and overall assessment.

Summarization is a cognitive process that condenses text into a shorter, more concise form while preserving its essential meaning. It requires readers to identify the main ideas, supporting details, and overall text structure. By summarizing, readers can improve their understanding of the text, as well as retain and recall important information. Hidi et al. [5] showed that summarization also helps learners identify gaps in their comprehension, promoting critical thinking and reinforcing information retention.

For summarization, we used the same dataset as Botarleanu et al. [1] in which summaries were annotated on multiple dimensions to capture different components reflective of their quality: main idea presence, detailing level, cohesion, objective language, wording, language beyond source text, and summary length.

Paraphrasing involves rewording or restating a given text while preserving its original meaning. It is a key skill in both language learning and cognitive development, as it enables individuals to process and internalize information by expressing it in their own words. McNamara et al. [14] argue that this active engagement with the material not only helps clarify understanding but also deepens comprehension by forcing the reader to focus on the core concepts and reorganize them in a new form.

We considered the dataset constructed by McCarty et al. [10] containing paraphrases written by students for target sentences. Expert annotators scored these paraphrases on 10 dimensions: garbage content presence, frozen expressions, irrelevancy, elaboration, semantic completeness, entailment, syntactic similarity, lexical similarity, paraphrase quality, and writing quality. The grading was initially performed on a 1-6 scale; however, the authors propose multiple options for splitting the grades into buckets: keeping the 1-6 scale, splitting the scale in three, and splitting the scale in two. We split the grades into three buckets, arguing that this approach preserves the necessary scoring information while reducing subjective bias.

While taking into account the size and split of the considered datasets, the self-explanation dataset was divided into 6k examples for training, 3k for testing, and 3k for validation. The think-aloud dataset was split into 7k for training, 2k for testing, and 2k for validation. For the summaries dataset, the split was 7k for training, 0.4k for testing, and 0.4k for validation. Lastly, the paraphrases dataset was divided into 1k for training, 0.3k for testing, and 0.6k for validation. It is important to note that no support-text contamination occurred between train/test/validation data for all learning strategies with supporting text.

3.2 LLM Fine-tuning

We explored fine-tuning open-source LLMs to enhance their performance on various scoring tasks corresponding to different learning strategies. Our goal was to improve the models' adaptability to the specific requirements of these educational evaluations. Through these experiments, we aimed to identify the optimal combination of fine-tuning strategies and model sizes for scoring tasks with limited annotated data.

We experimented with two training approaches for the selected LLMs:

Multi-task Training: In these experiments, we simultaneously fine-tune the models across all scoring tasks. We aim to test the hypothesis that training on a diverse set

Prompt content
Generated content

Rate the quality of the following performed task, based on the scoring rubric.

Task description:

Self-explanation	Think-aloud	Summarization	Paraphrasing
Explain the meaning of the text, elaborating beyond your initial understanding of the text. Context: <text> Target sentence: <target_sentence>	Report your thoughts that immediately come to mind regarding how you understand the meaning of the text. Context: <text> Target sentence: <target_sentence>	Summarize the following text. Context: <text>	Paraphrase the following sentence. Target sentence: <target_sentence>

Execution: <student_response>

Scoring rubric:
- <rubric_name_1>:
- - <score_1>: <score_description>
- - <score_2>: <score_description>
...
- - <score_m>: <score_description>
- <rubric_name_2>:
...
- <rubric_name_n>:
...

Response:
- <rubric_name_1>: <predicted_score>
- <rubric_name_2>: <predicted_score>
...
- <rubric_name_n>: <predicted_score>

Figure 1: Prompt structure.

of tasks enables the models to transfer knowledge between tasks, thereby enhancing their overall scoring capabilities. This approach also seeks to improve task performance with limited training data by leveraging insights from more data-rich tasks. For this, we append all training examples and randomize their occurrence. In this way, the models are trained on a mixture of scoring prompts without the risk of catastrophic forgetting [9]. This setup requires only a single model, so it is more efficient in terms of memory.

Single-task Training: In these experiments, we individually fine-tune models for each scoring task. We aim to test the hypothesis that specialized fine-tuning optimizes the model to perform at the highest capacity in the specific task. For this purpose, we fine-tune 4 independent models, all originating from the same reference model, each dedicated to one of the learning strategies: self-explanation, think-aloud, summarization, and paraphrasing. This setup expands the memory required to 4x the previous approach when deployed in the same application.

Besides the two previous approaches, proper prompt formatting is crucial for both optimizing the model's performance and ensuring fairness in comparisons. To maintain consistency, we apply the same prompt structure across all scoring tasks, with the only variations being the specific task descriptions and materials provided.

Our prompt format is illustrated in Figure 1. During training, the model receives the input highlighted in yellow, while gradients are computed based on the response section, shown in blue. Back-propagation is then applied to these responses to fine-tune the model's ability to generate accurate outputs. During inference, the model is given the input depicted in yellow and is expected to generate the appropriate response autonomously. In order to take advantage of the LLMs' understanding of language and its better assessment, we transform the datasets' scores from numeric format to text format (e.g., *1 - Poor, 4 - Excellent*). In this manner, the models can better associate the level of student performance with an actual score translation.

A detailed correspondence table and more prompt examples for each task can be found in the code repository. Moreover, to further enhance the description and model's understanding of the grading system, for each scoring rubric and each particular score value, we also append in the prompt a description of the student response (e.g., *Elaboration - Good - There is a response regarding the theme of the target sentence rather than a restatement of the sentence.*). These descriptions are derived from the instructions given to the annotators of each dataset.

3.3 Experimental Setup

We conducted experiments using two families of LLMs to ensure the stability of the proposed training methods. For the encoder-decoder architecture, we employed the Flan-T5 family [3], an instruction-tuned variant of the T5 model [18]. For the decoder-only architecture, we selected the instruction-tuned Llama 3.2 model [4]. We selected these two types of model architecture (i.e., encoder-decoder and decoder-only) due to their strong performance in instruction-following and generalization. Additionally, the two model families (i.e., Flan T5 and Llama) serve as representative examples of their respective architectures and their effectiveness across various tasks.

To maintain size consistency across both model families, we experimented with the following variants: Flan-T5 large (1B parameters), Flan-T5 xl (3B parameters), Llama 3.2 1B, and Llama 3.2 3B. These models are designed to be versatile and capable of addressing a wide range of natural language processing tasks, including deployment on resource-constrained devices.

We fully fine-tuned the models by updating their weights in full precision on an A100 80GB GPU. The training setup used for all our experiments is the following: constant learning rate of 1e-5 for Flan-T5 and 5e-6 for Llama 3.2, AdamW 8bit optimizer, and a batch size of 64 obtained by gradient accumulation.

4. RESULTS

We computed the results from Table 1 in terms of weighted F1 on the test partition of the datasets for each task and scoring dimension. We evaluated both sizes of the models (i.e., 1B and 3B) and both training setups (single-task - STL and multi-task - MTL). Moreover, we introduce a strong baseline with a larger model, namely zero-shot inferences using a Llama 3.3-70B-Instruct model, quantized in 4-bit. The weighted F1 is computed for an exact match between the generated and annotated ones. We also computed, for each task, an average across all rubrics to aggregate the scores and highlight the best-performing model.

For clarity and ease of interpretation, the highest scores within each scoring rubric are highlighted in bold, considering all model architectures, sizes, and training configurations. Additionally, to facilitate the comparison between the two training setups, we apply a color-coding scheme for each

Table 1: Performance results (F1 scores for STL and MTL setups on all 4 considered models, alongside the zero-shot setup).

	Flan T5 - 1B		Flan T5 - 3B		Llama 3.2 - 1B		Llama 3.2 3B		Llama 3.3 70B
	STL	MTL	STL	MTL	STL	MTL	STL	MTL	0-shot
	Self-explanation								
Paraphrase presence	0.71	0.71	0.71	0.74	0.71	0.77	0.78	**0.79**	0.10
Lexical change	0.35	0.35	0.35	0.37	**0.54**	**0.54**	0.53	**0.54**	**0.54**
Syntactic change	0.18	0.18	0.18	0.2	0.49	0.45	0.5	0.45	**0.53**
Misconception	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.76
Monitoring	0.95	0.95	0.95	0.95	0.95	0.97	0.97	0.97	**0.98**
Bridge presence	0.17	0.17	0.17	0.32	0.36	0.51	0.5	**0.52**	0.25
Bridge contribution	0.24	0.14	0.15	0.3	0.35	0.5	0.52	**0.53**	0.33
Elaboration presence	0.65	0.65	0.65	0.65	0.65	0.67	0.65	**0.7**	0.62
Life event	0.65	0.65	0.65	0.66	0.65	0.67	0.65	**0.7**	0.09
Overall	0.15	0.31	0.31	0.49	0.5	0.66	0.62	**0.67**	0.09
Averaged Rubrics	0.50	0.51	0.51	0.57	0.62	0.67	0.67	**0.69**	0.43
	Think-aloud								
Paraphrase presence	0.09	0.09	0.09	0.47	0.11	0.47	0.15	**0.48**	0.19
Lexical change	0.09	0.09	0.09	0.33	0.11	**0.35**	0.13	0.34	**0.35**
Syntactic change	0.09	0.09	0.09	0.2	0.1	0.23	0.12	0.22	**0.46**
Misconception	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	0.54
Monitoring	0.85	0.85	0.85	0.88	0.85	0.91	**0.92**	0.91	**0.92**
Bridge presence	0.23	0.23	0.23	0.35	0.23	0.44	0.34	**0.44**	0.20
Bridge contribution	0.27	0.27	0.27	0.37	0.31	0.46	0.34	**0.48**	0.26
Elaboration presence	0.4	0.4	0.4	0.42	0.51	0.51	**0.6**	0.57	0.46
Life event	0.4	0.4	0.4	0.45	0.64	0.55	**0.74**	0.63	0.27
Overall	0.34	0.34	0.34	0.56	0.56	0.65	0.57	**0.65**	0.10
Averaged Rubrics	0.37	0.37	0.37	0.50	0.44	0.55	0.49	**0.57**	0.37
	Summary								
Main Idea	0.52	0.52	0.69	0.78	0.76	**0.83**	0.75	0.8	0.30
Details	0.4	0.4	0.51	0.71	0.71	**0.8**	0.77	0.79	0.32
Cohesion	0.58	0.58	0.64	0.78	0.73	0.79	0.76	**0.8**	0.19
Objective language	0.65	0.65	0.68	0.69	0.71	0.73	0.75	**0.8**	0.63
Wording	0.35	0.35	0.51	0.61	0.62	0.7	0.76	**0.83**	0.41
Language beyond source text	0.31	0.31	0.71	0.79	0.68	0.76	0.82	**0.83**	0.47
Summary Length	0.38	0.38	0.51	0.63	0.53	**0.72**	0.67	0.71	0.28
Averaged Rubrics	0.45	0.45	0.60	0.71	0.67	0.76	0.75	**0.79**	0.37
	Paraphrasing								
Garbage Content	0.95	0.95	0.95	0.95	0.95	0.99	0.98	**0.99**	0.72
Frozen Expressions	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.76
Irrelevant	0.92	0.92	0.92	0.92	0.92	**0.95**	0.92	0.94	0.42
Elaboration	0.96	0.96	0.96	0.96	0.96	0.96	0.96	**0.97**	0.95
Semantic Completeness	0.41	0.41	0.42	0.67	0.14	**0.72**	0.63	**0.72**	0.24
Entailment	0.5	0.5	0.5	0.75	0.13	0.74	0.64	**0.76**	0.23
Syntactic Similarity	0.42	0.03	0.43	0.41	0.42	0.56	0.44	**0.67**	0.26
Lexical Similarity	0.21	0.21	0.21	0.61	0.05	0.62	0.37	**0.68**	0.35
Paraphrase Quality	0.19	0.26	0.19	**0.55**	0.17	**0.55**	0.43	0.51	0.18
Writing Quality	0.46	0.46	0.46	0.57	0.46	0.6	0.49	**0.66**	0.03
Averaged Rubrics	0.60	0.57	0.60	0.74	0.52	0.77	0.69	**0.79**	0.41

individual model size and type: green indicates cases where the multi-task learning (MTL) score exceeds the single-task learning (STL) score, red denotes instances where the STL score is higher, and uncolored cells represent cases where both setups yield identical scores.

5. DISCUSSION

Table 1 showcases that the multi-task learning setup consistently outperforms the single-task approach, achieving superior or comparable results across the majority of scoring rubrics. This finding suggests that training on a diverse set of tasks enhances the model's capability to generalize to unseen data, improves its understanding of the overall scoring framework, and mitigates the risk of overfitting. The observed trend is consistent across various model families and sizes, supporting the generalizability of this conclusion. The effect is more pronounced in larger models and newer architectures, whereas for Flan-T5 1B, the model exhibits underfitting due to its size and pre-instruction tuning, leading to negligible performance differences in the multi-task setting. These results highlight the benefits of multi-task training, as it not only enables better generalization in data-scarce sce-

narios but also improves performance across all tasks. This approach allows a single model to be efficiently deployed for scoring student responses across multiple tasks, thereby reducing memory consumption and computational costs.

An important observation is the superior performance of the Llama 3.2 3B model across all averaged rubric scores per task, arguing for its effectiveness for this application while remaining feasible for deployment on consumer-grade GPUs with low-latency inference. This advantage is attributed to its model size, pre-training corpus, and instruction-tuning data, which serve as a strong foundation for fine-tuning on specific scoring tasks. Moreover, our results indicate that our comparatively smaller models (1B and 3B), when fine-tuned in a multi-task setting, consistently outperform the larger zero-shot model (70B). This underscores that performance is not solely determined by model size but is significantly influenced by effective task-specific adaptation. Achieving high accuracy with a relatively compact model underscores the potential for scalable, automated scoring systems in diverse settings, enabling AI-driven assessment solutions without requiring high-end computational infrastructure.

The studies presented in the Related Work section report results that are not directly comparable to our approach. Nicula et al. [15, 16] reported results for the tasks of self-explanation and paraphrasing. However, due to class imbalances, the authors aggregated certain classes within the scoring rubrics, thereby simplifying the task. In contrast, our approach does not employ such modifications. For the self-explanation task, they achieved F1 scores between 0.72 and 0.89 across different evaluation criteria, including paraphrase presence, bridging presence, elaboration presence, and an overall score. In the paraphrasing task, they obtained F1 scores ranging from 0.68 to 0.86 for lexical, semantic, and syntactic similarity, as well as for overall paraphrase quality. Additionally, Botarleanu et al. [2] tackled the summary evaluation task as a regression problem rather than a classification task, reporting an overall R^2 of 0.64.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we explored the use of Large Language Models to automate the scoring of student responses to learning-strategy prompts, focusing on tasks such as self-explanation, think-alouds, summarization, and paraphrasing. Our approach aimed to address the challenges of scaling manual evaluation by leveraging fine-tuned LLMs trained on datasets annotated by expert educators. Through extensive experiments with both multi-task and single-task training setups and using models of different sizes and architectures, we investigated the extent to which LLMs replicate human scoring across multiple rubrics. One of the key findings of this work is the clear advantage of multi-task training. Models fine-tuned across a mixture of scoring tasks consistently outperformed those trained on single tasks. This result highlights the potential of transfer learning - the models generalize more effectively to unseen texts by training on diverse tasks, even in scenarios where task-specific data is scarce.

Moreover, our results indicate that smaller models (1B and 3B), when fine-tuned in a multi-task setting, consistently outperform a significantly larger zero-shot model (70B). This finding suggests that multi-task finetuning can be an efficient strategy for improving model performance without the computational overhead associated with scaling to larger architectures. The results indicate that task-specific adaptation fosters more nuanced understanding and contextual alignment compared to zero-shot approaches, where the model must rely solely on its pretraining distribution. This aligns with prior findings in transfer learning, emphasizing that targeted finetuning can bridge the gap between general-purpose knowledge and domain-specific expertise.

Furthermore, our work contributes to open education by using and fine-tuning open-source LLMs, promoting transparency and reproducibility. We open-sourced the models, training scripts, and evaluation frameworks, ensuring the research community can build on our work and adapt it to different educational contexts. These findings support that AI-driven assessment tools can efficiently support educators and learners in providing timely, personalized scoring on learning strategies, fostering better educational outcomes at scale.

Future work targets leveraging the advanced capabilities of LLMs to generate synthetic data that mimics the structure and content of diverse student responses. Combining this synthetic data with existing datasets through semi-supervised learning techniques could significantly enhance the model's capability to perform scoring tasks by providing a richer and more diverse training set. In semi-supervised learning, the synthetic data generated by larger models could serve as the unlabeled portion, which, when paired with smaller amounts of high-quality, human-annotated data, would help the model learn from both the expert annotations and the broader patterns in the synthetic examples. This hybrid approach has the potential to improve model generalization and prevent overfitting on small datasets. Moreover, chat models can be prompted to produce responses with varying levels of quality, emulating different student performances, thus creating a more robust training set for evaluating diverse learning strategies.

Additionally, future experiments will explore how the models perform on new, unseen tasks to evaluate their generalizability beyond the specific scoring tasks they were trained on. One direction is to test the models on essay scoring, which involves more complex and structured writing. By introducing new tasks, we can assess whether the models adapt to different forms of student work, providing further insights into their generalization capabilities.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University, by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 334906, and the grant of the Academy of Romanian Scientists, AOSR-TEAMS-IV Edition 2025-2026 "Digital Transformation in Science". The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

7. REFERENCES

[1] R.-M. Botarleanu, M. Dascalu, L. K. Allen, S. A. Crossley, and D. S. McNamara. Automated summary scoring with readerbench. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*, pages 321–332. Springer, 2021.

[2] R.-M. Botarleanu, M. Dascalu, L. K. Allen, S. A. Crossley, and D. S. McNamara. Multitask summary scoring with longformers. In *International Conference on Artificial Intelligence in Education*, pages 756–761. Springer, 2022.

[3] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[4] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] S. Hidi and V. Anderson. Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, 56(4):473–493, 1986.

[6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[7] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[8] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

[9] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

[10] P. M. McCarthy and D. S. McNamara. The user-language paraphrase corpus. In *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, pages 73–89. IGI Global, 2012.

[11] D. S. McNamara. Sert: Self-explanation reading training. *Discourse processes*, 38(1):1–30, 2004.

[12] D. S. McNamara. Self-explanation and reading strategy training (sert) improves low-knowledge students' science course performance. *Discourse Processes*, 54(7):479–492, 2017.

[13] D. S. McNamara, N. Newton, K. Christhilf, K. S. McCarthy, J. P. Magliano, and L. K. Allen. Anchoring your bridge: The importance of paraphrasing to inference making in self-explanations. *Discourse Processes*, 60(4-5):337–362, 2023.

[14] D. S. McNamara and J. L. Scott. Training reading strategies. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*, pages 387–392. Psychology Press, 2020.

[15] B. Nicula, M. Dascalu, T. Arner, R. Balyan, and D. S. McNamara. Automated assessment of comprehension strategies from self-explanations using llms. *Information*, 14(10):567, 2023.

[16] B. Nicula, M. Dascalu, N. N. Newton, E. Orcutt, and D. S. McNamara. Automated paraphrase quality assessment using language models and transfer learning. *Computers*, 10(12):166, 2021.

[17] E. Page. The imminence of grading essays by computer. *Phi Delta Kappan*, 47:238–243, 1966.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[19] S. Ruseti, M. Dascalu, A. M. Johnson, D. S. McNamara, R. Balyan, K. S. McCarthy, and S. Trausan-Matu. Scoring summaries using recurrent neural networks. In *Intelligent Tutoring Systems: 14th International Conference, ITS 2018, Montreal, QC, Canada, June 11–15, 2018, Proceedings 14*, pages 191–201. Springer, 2018.

[20] Y. Song, Q. Zhu, H. Wang, and Q. Zheng. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 2024.

[21] J. Wang and J. Liu. T-mes: Trait-aware mix-of-experts representation learning for multi-trait essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236, 2025.