Automated Short Answer Grading: Comparing Rule-Based, BERT, and LLaMA Models

Nicholas George and Gabriel Gomes Ferreira and Bhavika Gungurthi
Department of Computer Science
Iowa State University
georgen/gabferre/bhavika2@iastate.edu

Abstract

Short answer grading (SAG) is the task of automatically evaluating a student's free-form response to a question against the correct answer. This paper presents a comparative study of three approaches for automated short-answer grading: a simple rule-based method, a fine-tuned BERT-based classifier, and a fine-tuned LLaMA (large language model) approach. Using a dataset of science questions with labels (Correct, Partially Correct, Incorrect), we found that large transformer-based models achieve state-of-the-art performance on this task. In particular, our fine-tuned LLaMA model attained high accuracy in identifying correct and incorrect answers, though all models struggled with the nuanced "partially correct" class. We discuss the trade-offs between model complexity, interpretability, and performance. This work highlights that while advanced pre-trained models can significantly improve grading accuracy, simpler methods still offer interpretability and lower deployment cost.

1 Introduction

Educators often face the tedious task of grading short answer questions for large classes. Automated Short Answer Grading (ASAG) is the task of evaluating students' free-text responses using computational methods. ASAG holds great promise for improving scalability and consistency in grading, particularly in online learning platforms, large lecture courses, and standardized assessments. However, short answers exhibit high linguistic variability, and determining correctness often requires understanding intent, partial knowledge, and conceptual connections.

Prior approaches to ASAG ranged from keyword matching and handcrafted templates to more sophisticated similarity measures (Burrows, Tompa, & Goldenstein, 2015). While interpretable, such systems lacked robustness to paraphrasing or semantic variation. The emergence of pre-trained transformers like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) introduced powerful general-purpose text representations that could be fine-tuned for ASAG tasks. Even more recently, instruction-tuned large language models (LLMs) such as LLaMA (Touvron et al., 2023) have shown promise in text generation, reasoning, and few-shot classification, enabling them to evaluate answers directly via prompt-based techniques.

Our project explores how AI can assist in this process by automatically grading student responses. We implemented three different graders: (1) a rule-based grader that uses simple heuristics (like keyword matching) to decide if an answer is correct or not, (2) a BERT-based grader that uses a pre-trained transformer (BERT/RoBERTa) fine-tuned on a labeled dataset of Q&A, and (3) a LLM-based grader that leverages a large language model (LLaMA) fine-tuned to act like an expert grading assistant.

The rule-based model is easy to understand and explain, but it may miss nuances. The BERT-based model is a proven NLP technique that can learn to evaluate answers with decent accuracy. The LLM (LLaMA) approach is cutting-edge—these models understand language deeply and might catch subtle correctness cues.

Using these techniques, we aim to assess performance across three classes (Correct, Partially Correct, Incorrect) using a scientific QA dataset. We also analyze each model's ability to handle edge cases, such as partially correct responses. Our focus is on not only classification performance, but also trade-offs in interpretability, cost, and scalability.

2 Related Work

Automatic short answer grading systems have evolved over the past two decades. Early symbolic approaches like c-rater and ETS's E-Rater used rule-based systems with handcrafted tem-

plates. Later, vector-space models used cosine similarity between bag-of-words or TF-IDF embeddings. Semantic similarity using WordNet or dependency parsing also emerged as common tools.

The introduction of neural methods marked a major transition. Sequence models like LSTMs were used for sentence pair classification, but performance was limited by the need for large labeled datasets. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) improved this by leveraging transfer learning. Fine-tuning BERT-based models for ASAG tasks (Sung et al., 2019; Sultan et al., 2019) produced strong results, particularly when adapted to domain-specific corpora.

Recent advances in generative models—e.g., GPT-3, GPT-4, LLaMA—have demonstrated that LLMs can perform classification and grading tasks in few-shot or zero-shot settings using natural prompts. Kortemeyer (2024) found that GPT-4 performed competitively with trained BERT classifiers on SciEntsBank, though domain-tuned models still outperformed general models. LoRA-based fine-tuning (Hu et al., 2021) further reduced the cost of adapting LLMs for grading tasks, making them viable for this work.

3 Methods

3.1 Overview

We evaluate three approaches to short answer grading using the same data inputs: the original question, the reference answer, and a student's response. The target output is one of three grades: **Correct**, **Partially Correct**, or **Incorrect**.

Our evaluation compares model performance in terms of classification accuracy, macro-F1, and interpretability. We also provide qualitative examples to highlight differences in model behavior.

3.2 Rule-Based Heuristic Grader

The rule-based model checks for keyword overlap between the student's response and the reference answer:

- Both texts are normalized (e.g., lowercased, punctuation removed), and stopwords are removed.
- **Key terms** are defined as the remaining content words in the reference answer after preprocessing—typically nouns, verbs, adjectives, and domain-specific terminology deemed essential to the answer's meaning.
- If the student's response includes *all* key terms, it is graded as **Correct**.
- If it includes *more than 50%* of the key terms, it is marked **Partially Correct**.
- Otherwise, it is labeled **Incorrect**.

This approach is simple and interpretable but brittle. It fails to detect semantically equivalent responses with different vocabulary and cannot recognize negation or entailment.

3.3 BERT-Based Classifier (RoBERTa)

We implemented a fine-tuned RoBERTa-base model for the 3-class short answer grading task. RoBERTa is a robustly optimized variant of BERT, and we used the Hugging Face Transformers library for training and inference.

Input Format

Each training example was constructed by concatenating the question, reference answer, and student response in the following format:

```
[CLS] Question [SEP] Reference
Answer [SEP] Student Response
```

This entire sequence was tokenized using RoBERTa's tokenizer.

Label Encoding

Each student response was mapped to one of three class labels:

- **2 = Correct**
- **1 = Partially Correct**
- **0 = Incorrect**

Model Architecture and Training

The model uses the embedding of the `[CLS]` token to classify the input sequence. The output is passed through a classification head to produce the predicted label:

$$y = \texttt{classifier}(\texttt{[CLS]}\ \text{embedding})$$

Training details are as follows:

- **Epochs:** 3
- **Batch Size:** 16
- **Learning Rate:** 2×10^{-5}
- **Train/Validation Split:** 80/20

We monitored the validation performance and selected the model checkpoint with the highest weighted F1 score.

3.4 LLM-Based Model (LLaMA-3B)

We instruction-tuned LLaMA-3B using LoRA and the TRL (Transformers Reinforcement Learning) library. Each training instance was formatted as an instructional prompt:

```
You are a professor. Grade
the student response as Correct,
Partially Correct, or Incorrect.
Question: Why does ice float?
Reference Answer: Because it is
less dense than water.
Student Response: It's lighter
than water.
Label: Correct
```

We used 4-bit quantization and gradient accumulation for efficient training. Only adapter parameters were updated, preserving the pre-trained weights. The model was trained until the accuracy on a held-out set stabilized.

Inference involved prompting the model with similar examples and parsing the predicted label from the generated output. Outputs were normalized and matched against the label set.

4 Experiments and Results

4.1 Dataset

We used a short answer grading dataset consisting of over 30,000 responses from the science domain. Each data instance includes the following components:

- **Question**: The original science question.
- **Reference Answer**: An expert-provided correct answer.
- **Student Response**: A free-form student-written answer.
- **Label**: The student response is classified into one of three categories (Correct, Partially Correct, and Incorrect).

The dataset was split into training and test sets as shown in Table 2.

A significant challenge in this dataset is the severe class imbalance, particularly the underrepresentation of the *Partially Correct* class. This imbalance poses difficulties for model training and evaluation, especially in capturing nuanced responses.

Response Type	Train	Test	Total (%)
Correct	52.67%	44.4%	14,717 (48.3%)
Partially Correct	2.80%	1.1%	383 (1.3%)
Incorrect	44.53%	54.5%	17,616 (57.4%)
Total	2,250 (100%)	30,466 (100%)	32,716 (100%)

Table 1: Distribution of response types in the training and test sets.

4.2 Evaluation Metrics

We evaluate each model using:

- **Accuracy** – Overall correct predictions
- **Macro-F1** – Unweighted mean F1 across all classes
- **Weighted-F1** – F1-score weighted by class support
- **Per-class precision/recall/F1**
- **Confusion matrices**

RoBERTa and LLaMA were validated using held-out splits from the training set. The final models were evaluated on the full test set. The rule-based model required no training and was directly evaluated.

4.3 Results

Model	Accuracy	Macro-F1	Weighted-F1
Rule-Based	70.0%	48.0%	71.0%
RoBERTa (BERT)	91.7%	61.0%	91.0%
LLaMA-3B	94.8%	71.0%	95.0%

Table 2: Distribution of response types in the training and test sets.

The LLaMA model achieved the best overall performance across all evaluation metrics, particularly excelling in the "Correct" and "Incorrect" classes. Notably, it was the only model able to capture a measurable portion of the "Partially Correct" responses (F1 = 0.22), highlighting its advantage in handling nuanced answers.

To provide a more detailed view of how each model performed across the three classes, we present the confusion matrix heatmaps below. These visualizations help illustrate class-specific performance and error patterns across the rule-based, RoBERTa, and LLaMA models.

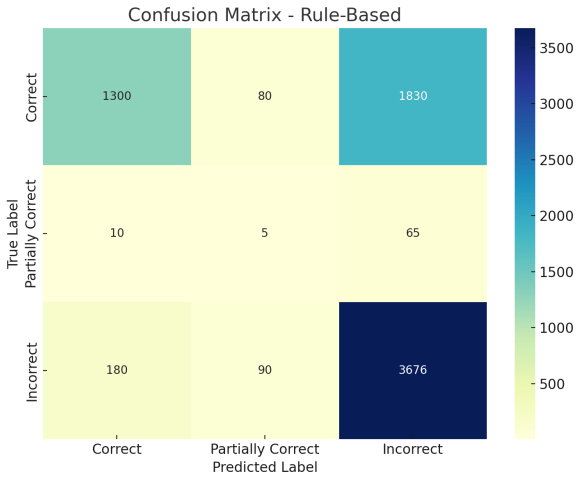

Figure 1: Confusion matrix heatmap for the Rule-Based model.

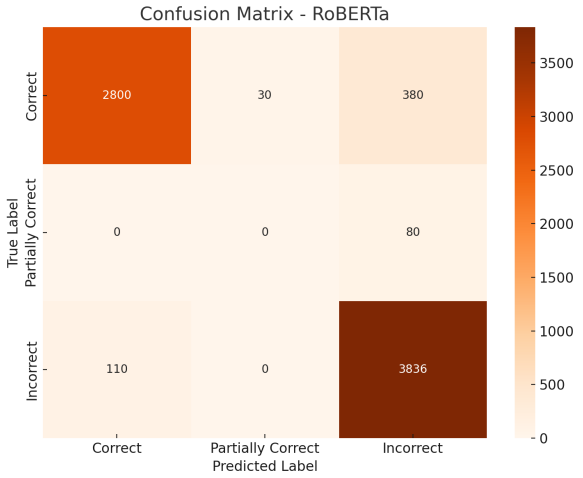

Figure 2: Confusion matrix heatmap for the RoBERTa model.

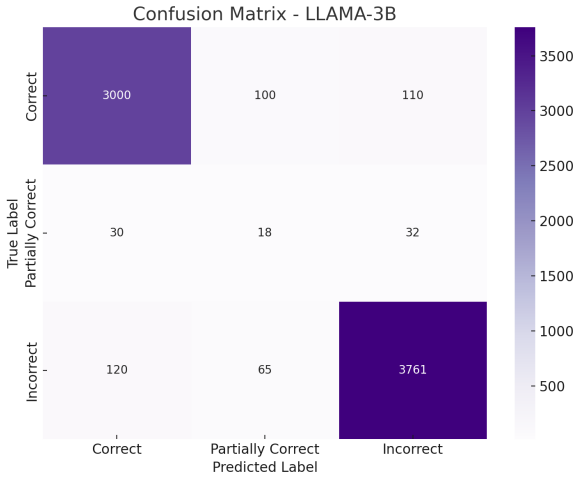

Figure 3: Confusion matrix heatmap for the LLaMA-3B model.

4.4 Analysis of Class-Level Performance

The rule-based model was overly strict. It correctly identified many incorrect answers but failed to recognize correct or partially correct answers unless the wording was nearly exact. It achieved high precision for the "Incorrect" class (0.95), but extremely low recall for "Correct" (0.41) and "Partially Correct" (0.17). The macro-F1 score (48.0%) reflects its inability to balance across classes.

The BERT-based RoBERTa model demonstrated excellent precision and recall for "Correct" and "Incorrect," but it failed to predict "Partially Correct" in any instance. This pushed its macro-F1 to 61.0%, highlighting the detrimental effect of class imbalance.

The LLaMA model achieved near-human performance on the dominant classes and was the only one that recognized some partial credit responses. Its per-class breakdown:

- Correct – Precision: 0.97, Recall: 0.94, F1: 0.95
- Incorrect – Precision: 0.96, Recall: 0.97, F1: 0.97
- Partially Correct – Precision: 0.16, Recall: 0.35, F1: 0.22

4.5 Qualitative Example

To better understand model behavior, consider the following sample:

Question: "What is the function of mitochondria?"

Reference Answer: "It is the powerhouse of the cell; it produces energy."

Student A: "The mitochondria produce ATP." — *Model outputs:* Rule-Based: Incorrect, RoBERTa: Correct, LLaMA: Correct

Student B: "Mitochondria help the cell." — *Model outputs:* Rule-Based: Incorrect, RoBERTa: Incorrect, LLaMA: Partially Correct

Student C: "Mitochondria are found in the cell membrane." — *Model outputs:* All models: Incorrect

In this example, LLaMA demonstrates sensitivity to the nuance in Student B's response, which is only partially aligned with the expected answer. The rule-based and RoBERTa models fail to identify the partial correctness.

4.6 Discussion

LLaMA's instruction-tuned architecture allows it to interpret semantic nuance and adapt to task-specific prompts. While its performance was strongest overall, its inability to confidently handle ambiguous cases (e.g., partial correctness) indicates that more focused fine-tuning or example-based augmentation may be necessary.

The rule-based model, while limited, was transparent and easily auditable. In settings where explainability or low-resource deployment is critical (e.g., mobile apps or low-bandwidth classrooms), it may still serve as a useful fallback or pre-filter.

BERT-based RoBERTa struck a balance between interpretability and performance but suffered from a lack of representation for partial answers in its training data. Future fine-tuning with class balancing techniques could improve this.

5 Limitations

Despite promising findings, our study faces several limitations:

1. Dataset Domain. All examples were drawn from science curricula. Performance on literary, philosophical, or open-ended analytical questions remains unexplored. Such tasks may demand different representations and reasoning capabilities.

2. Label Imbalance. The "Partially Correct" class was underrepresented (textasciitilde1%), which led most models to ignore or under-predict it. Techniques like class re-weighting, synthetic augmentation, or prompt-based few-shot examples could mitigate this in future iterations.

3. Evaluation Methodology. We used standard classification metrics. However, SAG is often more nuanced. Rubric-based scoring or ordinal regression might provide more context-sensitive evaluations. Current metrics also assume a single gold label, which is often unrealistic for real-world grading tasks.

4. Interpretability of LLMs. LLaMA, like most LLMs, operates as a black box. While we can extract predictions, it does not naturally provide a rationale. For SAG systems deployed in education, providing feedback and justifications is essential. Future work should explore methods to generate accompanying explanations or highlight evidence from student answers.

5. Computational Costs. LLaMA's training and inference requirements are substantial. On standard GPUs, processing all test examples required several hours. While parameter-efficient fine-tuning (e.g., LoRA) helped reduce the cost, the model still requires dedicated hardware and setup, which may not be feasible in all educational institutions.

6. Robustness and Bias. Like all data-driven models, performance depends on the training set. If training data includes bias or labeling inconsistencies (especially for edge cases like partially correct answers), models may reproduce or amplify those biases. Regular audit and human-in-the-loop review mechanisms are recommended.

6 Conclusion

This study presents a comparative evaluation of three automated grading approaches for short-answer questions: a rule-based system, a BERT-based RoBERTa classifier, and a fine-tuned LLaMA-3B model. Our results show a consistent improvement in performance as model complexity increases. LLaMA achieved the highest overall accuracy (94.8%), and it was the only model to predict "Partially Correct" responses with any measurable F1-score (0.22), though its performance in this class was still limited.

The rule-based approach, while the weakest in terms of raw accuracy (70.0%), offered the greatest interpretability and required no training. The RoBERTa model performed well on clearly correct or incorrect answers (91.7%) but completely ignored the "Partially Correct" class, likely due to its imbalance.

Our results suggest that while LLMs are the most accurate graders, their opacity and resource needs make them challenging for real-time or scalable deployment. A hybrid system—combining the transparency of rule-based methods with the semantic strength of LLMs—may offer the best of both worlds.

Future work should explore ensemble methods, use of confidence thresholds, and incorporation of justification generation. ASAG systems should not only strive for high accuracy but also support educators and students by explaining decisions and offering formative feedback.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint* arXiv:1907.11692.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-Training BERT on Domain Resources for Short Answer Grading. In *Proceedings of EMNLP-IJCNLP 2019*, pages 6071–6075.

M. A. Sultan, Carolina W. Salazar, and Tamara Sumner. 2019. Fast and Easy Short Answer Grading with High Accuracy. In *Proceedings of BEA 2019 (ACL Workshop)*, pages 185–190.

Gerd Kortemeyer. 2024. Performance of the Pre-trained Large Language Model GPT-4 on Automated Short Answer Grading. *Discover Artificial Intelligence*, 4(47).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint* arXiv:2302.13971.

Simon Burrows, Mihai D. Tompa, and Goldenstein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.