An-Najah National University

College of Engineering and Information Technology

Information Technology Department (IT)

Natural Language Processing
(10672352)

Automatic Essay Grading using Instruction-Tuned Transformers

Prepared by:
Bilal Anabosi
Toqa Asedah
Ahmad Istieteh

Date of Submission: May 2025

Abstract :

Manual essay grading is a labor-intensive process prone to inconsistency and bias. This project presents an automated essay grading system utilizing a 4-bit quantized instruction-tuned transformer model, specifically the *Mistral-7B-Instruct-v0.2*. The system is designed to assess student essays by referencing the original question, a model answer, and a predefined mark scheme. Using LoRA (Low-Rank Adaptation) and efficient fine-tuning techniques on limited computational resources, the model was trained to generate accurate scores (0–4) and provide rationale aligned with expert grading standards. Evaluation metrics indicate strong performance, with a Pearson correlation of 0.91 and a within-1 score accuracy of over 99%. Furthermore, the model demonstrates high semantic similarity between human and machine-generated rationales. This work highlights the potential of instruction-tuned LLMs for scalable, consistent, and interpretable assessment in educational settings.

Acknowledgements :

We would like to express our sincere gratitude to our supervisor, Dr. Hamed Abdelhaq, for his invaluable guidance, continuous support, and constructive feedback throughout the duration of this project. His expertise in Natural Language Processing and dedication to student success have been instrumental in shaping our research. We also extend our appreciation to An-Najah National University and the Department of Information Technology for providing the resources and environment necessary for completing this work. Finally, we are thankful to our families and peers for their encouragement and support throughout this journey

Contents

1. Introduction

Manual essay grading is time-consuming and often inconsistent. This project developed an automated grading system using a 4-bit quantized instruction-tuned language model, *Mistral-7BInstruct-v0.2*, to predict essay scores (0–4) and provide rationale for those scores. The system uses the essay question, a reference answer, the student's answer, and a mark scheme to simulate expert grading—offering consistency, scalability, and explainability.

2. Dataset Preparation

2.1 Source and Structure

The dataset included:

- Essay question
- Expert reference answer
- Student answer
- Mark scheme (0–4 points) sometimes more
- score (0–4)
- rationale

Data was stored in a structured JSONL format, and it was AI generated.

2.2 Instruction Formatting

Each entry was converted into a prompt designed to guide the model to act like an expert examiner, instructing it to assign a score based on the mark scheme and provide a rationale and these are some examples of the instructions put:

```
Instructions:
- Grade the student answer on a scale from 0 to 4 based strictly on the mark scheme. - For each
  criterion, assess whether it was satisfied.
- Provide a detailed and objective rationale explaining the score and make it related to the question and the
  answer contextually.
- Always provide a score of 4 even if you face more than 5 points in mark scheme try to combine everything
  and give a score of 4 the same goes when less than 4.
- Be concise, specific, and professional in your explanation."
```

2.3 Splitting and Tokenization

- Dataset split: 80% training (2,838), 10% validation (355), 10% test (355)
- Tokenization handled up to 2048 tokens

- Prompt-output pairs followed the format used by Unsloth and Mistral models

3. Model Fine-tuning

3.1 Model and Hardware

- Base model: "unsloth/mistral-7b-instruct-v0.2-bnb-4bit"
- Quantization: 4-bit (nf4), float16 compute.
- Training on Google Colab with Tesla T4 GPU

3.2 LoRA (Low-Rank Adaptation)

Used to fine-tune efficiently with limited resources. Final configuration:

```
model = FastLanguageModel.get_peft_model( model, r=32, lora_alpha=32, lora_dropout=0.0,
#cant do more because of unsloth target_modules=["q_proj", "k_proj", "v_proj", "o_proj",
"gate_proj", "up_proj", "down_proj"], bias="none", use_gradient_checkpointing=True,
random_state=42,
)
```

3.3 Training Setup

- Epochs: 1
- Effective batch size: 16
- Optimizer: AdamW (8-bit)
- Learning rate: 2e-4
- Losses: Training loss ≈ 0.28, Validation loss ≈ 0.34

3.4 LoRA Parameter Experiments

Tested various combinations:

- Ranks: 16, 32, 64
- Alpha: 32, 64, 128

Best setup: r=32, lora_alpha=32

- Balanced performance and efficiency
- Avoided overfitting
- Generalized well on unseen examples **Challenges with higher ranks (r=64):**

- Overfitting signs (verbose rationales, poor generalization)
- More memory usage and slower training Diminishing returns on small datasets

4. Model Evaluation

4.1 Setup and Inference

- Inference temperature: 0.1 (slightly varied outputs)
- Extracted score using regex
- Defaulted to 0 if no score found

4.2 Performance Metrics

Language Modeling

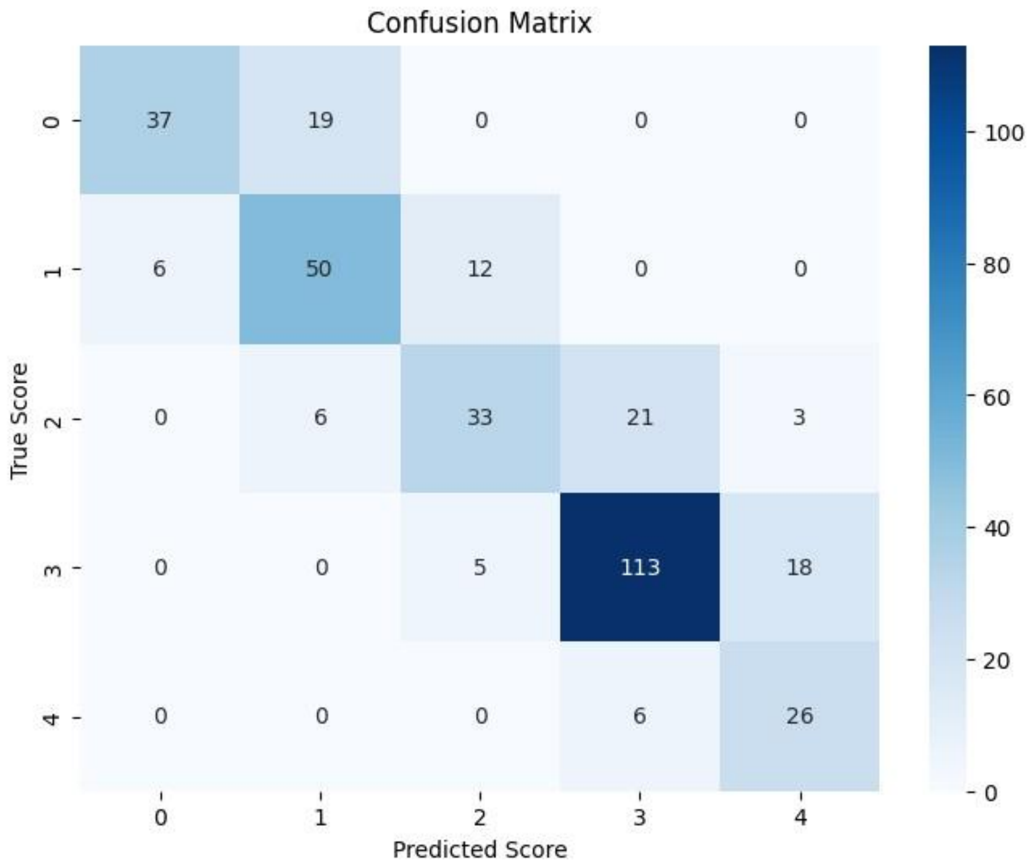
- **Perplexity**: 1.52 (indicates confident predictions)

Score Prediction Accuracy

- **Mean Absolute Error (MAE)**: 0.2789
- **Root Mean Square Error (RMSE)**: 0.5439
- **Pearson Correlation**: 0.9129 (strong linear correlation)
- **Exact Match Accuracy**: 72.96%
- **Within-1 Accuracy**: 99.15%

F1 Scores by Class

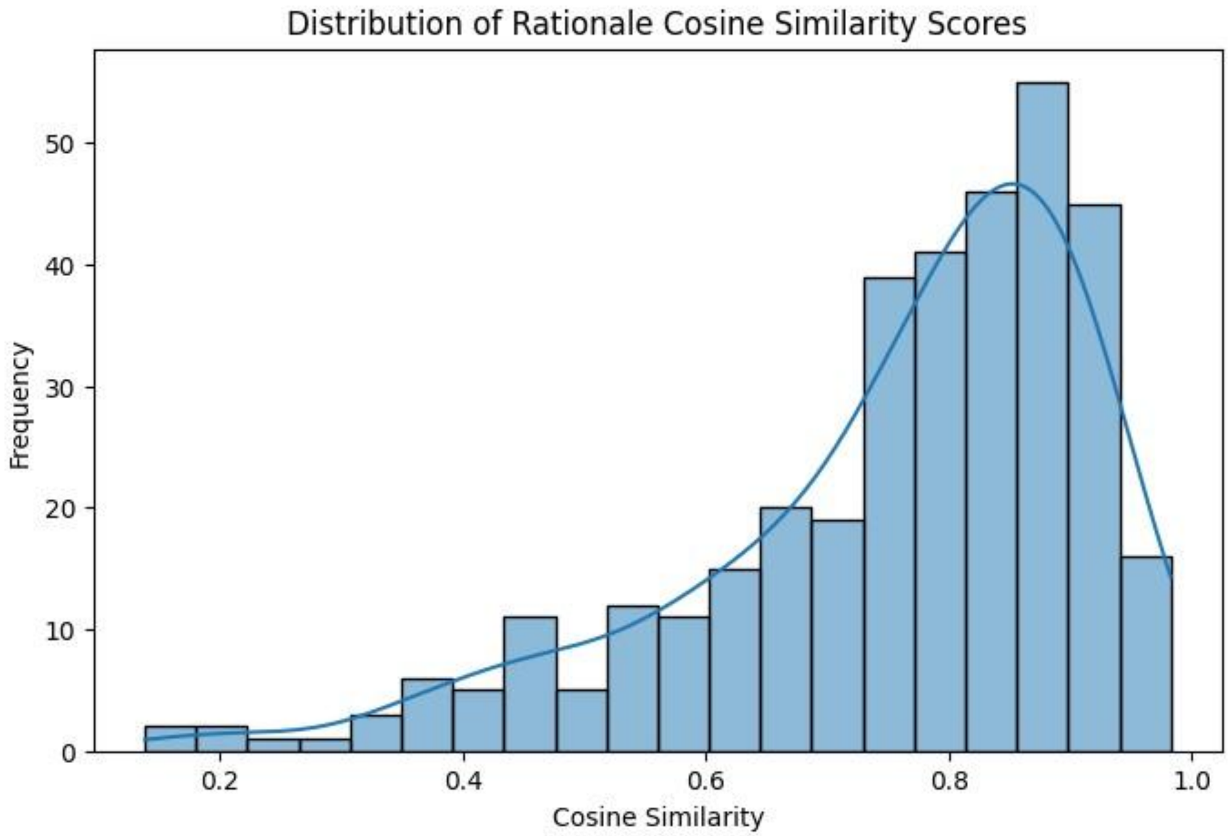
- Score 0: 0.75
- Score 1: 0.70
- Score 2: 0.58
- Score 3: 0.82
- Score 4: 0.66



Score 2 had the lowest F1, likely due to overlap with adjacent scores or fewer examples.

Rationale Similarity

- **Average Cosine Similarity** (model vs human rationale): 0.7561
- Indicates model-generated rationales are semantically close to human-written ones.



5. Results and Discussion

- The model effectively learned to replicate expert scoring behavior.
- **High exact match (73%)** and **near-perfect within-1 accuracy (99%)** show strong reliability.
- **Pearson correlation (0.91)** and **low MAE (0.28)** confirm predictive accuracy.
- **Rationale similarity (0.76)** supports interpretability.
- Best performance achieved with r=32, lora_alpha=32 — higher values led to overfitting or inefficiencies due to dataset size and limited compute resources.

6. Conclusion

This project successfully demonstrated that an instruction-tuned 4-bit *Mistral-7B* model can be fine-tuned to perform accurate and explainable essay grading, while still there is a margin of development using better datasets and better computational power. Key achievements include:

- Decent scoring accuracy

- Semantically meaningful rationale generation
- Efficient training using LoRA and 4-bit quantization
- Scalable grading system suitable for educational use