Interim Presentation

Automated Answer Correctness Evaluation

BY: Orel Cohen, Nathan Beer, Tomer Bengayev

Project Overview & Specifications

Problem Statement & Task Definition

- **Motivation:** Manual grading of open-ended CS questions is slow and inconsistent. We aim to automate correctness evaluation for student answers in data structures & algorithms.
- **Task:** Build a Regression model that assesses answer correctness using semantic understanding rather than keyword matching.

Evolution from Proposal

- **Refined Scope:** Shifted from broad classification to a more robust system resistant to keyword-based errors.
- **Improved Data Strategy:** Added Hard Negative Mining to handle imbalance and improve generalization; expanded synthetic data generation.
- **Defined Baselines:** Adopted Sentence-BERT and Zero-Shot GPT-4 for F1/RMSE comparisons.

Project Innovation & Contributions

- **Triple-Context Model:** End-to-end Transformer (BERT/RoBERTa) architecture capturing relations among *answer*, *question*, and *reference solution*.
- **Enhanced Dataset:** GPT-4 generated diverse synthetic examples to fill coverage gaps.
- **Modern Approach:** Fully neural, no manual features or old similarity metrics; focuses on deep semantic reasoning in CS content.

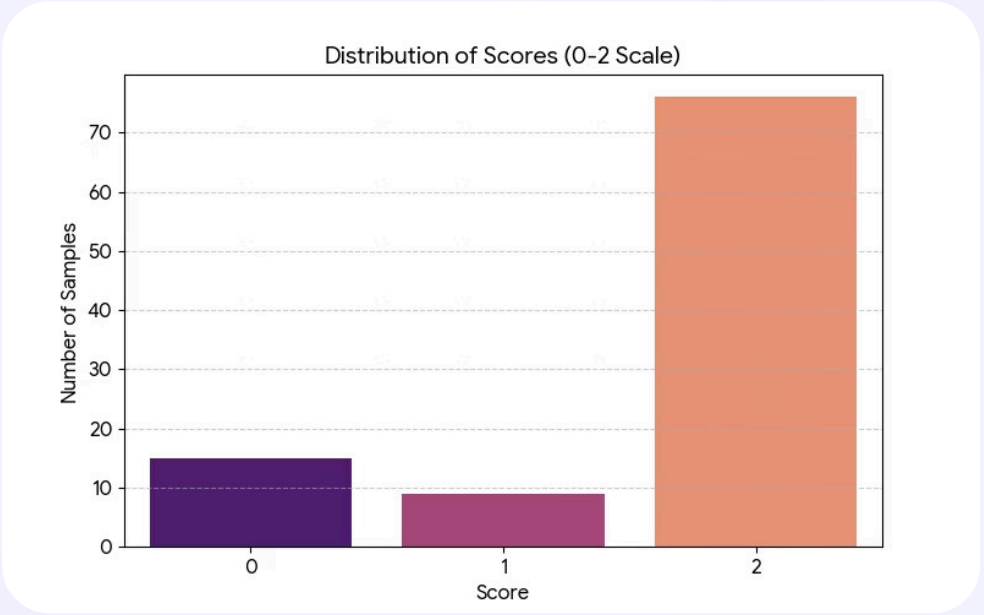
Previous work

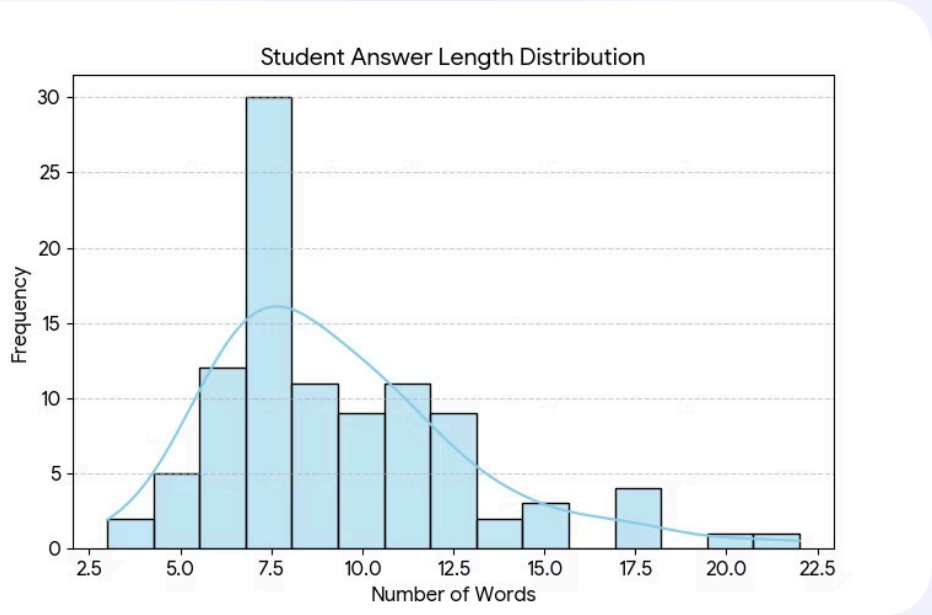
Title	Task	Method	Data	Results	Relation
NLP-based Automatic Answer Script Evaluation (2018)	Automatic scoring of student answers using text similarity	Word2Vec/FastText, Syntactic Analysis, or simple CNN/RNN networks	Sample of 30 descriptive questions and students answer to these questions.	F-score technique is greater than the bag-of-words based summarization technique.	Provides a similarity-based baseline for automatic answer grading
The Impact of Transformer Models and Regression Algorithms on Automated Short Answer Grading (2025)	Automatic grading of short answer questions (SAG) using continuous score prediction	Fine-tuned transformer model with Regression Head layer and MSE Loss	AMPLE Learning Management Software of AMRITA	BERT model with Random Forest algorithm gave the lowest MSE, RMSE, and the best R^2 score among all configurations	Justifies the choice of the Transformer and the use of regression as the ranking method
A review on student automatic grading system (2022)	Review and analyze existing automatic grading systems for essays and short answers	Survey prior AES research, datasets, feature-engineering approaches, and ML/NLP models	SRA / BEETLE corpus, SCIENTSBANK, ASAP, RTA, TOEFL11 corpus, ICLE, AAE	Other systems reviewed typically showed QWK scores in the range of **0.74 to 0.78** and a correlation coefficient of **0.76** between human and automated scores	Provides background on grading methods and similarity-based scoring

Dataset

- **Source:** Hybrid dataset that combines real data ("**MohlerASAG**") with synthetic data using a large language model (LLM) such as GPT-4.
- **Synthetic Data Purpose:** 100 synthetic questions and student answers were generated with GPT 4 .
  For each question, answers include correct, partially correct and incorrect responses.
- **Fields:** Question, Student answer (text), Correct answer (text), Score (0-2, 0=incorrect, 1=partial answer, 2=complete and correct answer).
- **Examples:** 7358 question and answers for science concepts (data structure/algorithms courses).
- **Label Collection:** Synthetic labels are generated via prompt design.
- **Hard Negative Mining:**Goal Challenge DeBERTa to identify logical connections and not just word similarities

```
estion,reference_answer,student_answer,score
What is the time complexity of Merge Sort?,"Merge Sort always has a time complexity of O(n log n) in all cases (best, average, worst).",Merge Sort is O(n^2) because it has nested loops during the merge proces
Explain the difference between a Process and a Thread.,"A process is an independent execution unit with its own memory space, while a thread is a subset of a process that shares memory with other threads.",A
What is a Primary Key in a Database?,"A primary key is a unique identifier for a record in a database table, and it cannot contain NULL values.",It is a key that allows duplicate values so we can group data.,
What is the purpose of a Hash Function?,"A hash function maps data of arbitrary size to fixed-size values, typically used for indexing in a hash table.",It creates a unique digital fingerprint for a piece of
What is the time complexity of Merge Sort?,"Merge Sort always has a time complexity of O(n log n) in all cases (best, average, worst).",The complexity is O(n log n) but it requires extra space O(n).,0.9
Explain the difference between a Process and a Thread.,"A process is an independent execution unit with its own memory space, while a thread is a subset of a process that shares memory with other threads.",Th
What is a Primary Key in a Database?,"A primary key is a unique identifier for a record in a database table, and it cannot contain NULL values.",A unique field that identifies each row in a table.,1.0
What is the purpose of a Hash Function?,"A hash function maps data of arbitrary size to fixed-size values, typically used for indexing in a hash table.",It creates a unique digital fingerprint for a piece of
hat is the time complexity of Merge Sort?,"Merge Sort always has a time complexity of O(n log n) in all cases (best, average, worst).",The complexity is O(n log n) but it requires extra space O(n).,0.9
```





Baseline Solution: Fine-tuned BERT Classifier

Model Architecture

Pretrained Model: SBERT

Method: Pre-trained off-the-shelf model without additional training on the project data (Zero-shot).

Technique: Calculation of Cosine Similarity between the vector (Embedding) of the student's answer and the vector of the reference answer. The score was converted to a scale of 0-2 for comparison.

Working assumption: Simple vector semantic similarity will be sufficient to assess the correctness of an answer.

Metric	Baseline (SBERT)	Our Model (DeBERTa)	Improvement
Pearson Correlation	0.1192	**0.7089**	**+495%**
RMSE (Error)	0.4526	**0.4042**	**-11%**

Baseline Performance & Error Analysis

Analysing the training process and common misclassifications provides valuable insights for future improvements.

Training Curves

Training and validation loss across epochs, indicating stable convergence with a small generalization gap.

Error Inspection: Examples & Observations

Lessons Learned

- Model struggles to distinguish between adjacent score categories (e.g., Score 1 vs. 2).
- Reliance on a single reference answer limits recognition of diverse correct formulations.

Next Steps

- Apply advanced data augmentation techniques to improve robustness.
- Explore multi-task learning for joint score and feedback prediction.

Example Misclassifications:

- Actual 2 → Predicted 1: Mostly correct answer missing a key concept, leading to underestimation.
- Actual 1 → Predicted 0: Relevant terms present, but explanation lacked coherence.

Project Roadmap: Advancing Automated Scoring

Our future work focuses on enhancing model accuracy and robustness through strategic data and model improvements.
Task ownership: Orel – data & augmentation, Nathan – modeling, Tomer – evaluation & reporting.

Task	Timeline	Goal/Outcome
Data Augmentation & Cleaning	Week 10	Expanded and more diverse training set; improved class balance and generalisation.
Explore Transfer Learning Models	Week 10	Quantitative comparison of RoBERTa/DeBERTa; selection of best-performing architecture.
Multi-task Learning Integration	Week 11	Joint learning of score and feedback tags; potentially more interpretable model.
Extended Error Analysis	Week 11	Detailed understanding of model limitations; actionable insights for fine-tuning.
Prepare Final Presentation	Week 12	Polished presentation summarising findings and next steps.

Technical Scope & Models

Data Augmentation Plan

Utilise back-translation, paraphrasing with LLMs, and noise injection to create varied training examples, for underrepresented classes.

Transfer Learning Options

Evaluate advanced Transformer models (RoBERTa, DeBERTa) for improved contextual understanding.

Error Analysis Plan

Fine-grained error categorization (factual gaps, coherence issues, missing components) to guide model refinement.