IMPROVING AUTOMATIC SHORT ANSWER GRADING WITH DEEP LEARNING ARCHITECTURES

A Synopsis

For the award of

Doctor of Philosophy

In

Computer/ IT Engineering

Submitted By:

Gamit Nikunj Chunilal
[179999913009]

Supervisor:

Dr. Ajay N. Upadhyaya
Professor, SAL Engineering & Technical Institute
SAL Education, Ahmedabad
Gujarat

Submitted to

Gujarat Technological University
Ahmedabad, Gujarat, India.

Supervisor	DPC Member 1	DPC Member 2
Dr. Ajay N. Upadhyaya	Dr. Amit Ganatra	Dr. Amit Thakkar

Contents

1. Abstract

Automatic Short Answer Grading (ASAG) plays a key role in scalable educational assessment by enabling computational evaluation of students' free-text responses with consistency and efficiency. Advances in Natural Language Processing, particularly transformer-based models such as BERT and Sentence-BERT, have improved meaning-based similarity estimation. However, these approaches often fail to capture true conceptual correctness, as high semantic similarity does not guarantee understanding—especially when key concepts are missing, answer length biases exist, or logical contradictions such as negation occur.

This research reframes ASAG as a structured educational reasoning problem rather than a pure similarity task. It addresses three core limitations: length discrepancy, where models struggle with varying answer lengths; correctness gap, where similarity does not reflect concept coverage; and Discrete Boundary Collapse (DBC), where continuous vector spaces fail to preserve clear grading distinctions.

To overcome these issues, a multi-architecture framework is proposed. First, a Universal Length-Adaptive Design routes answers to suitable sub-models based on length, integrating BM25, TF-IDF, SBERT, BERT, and Bi-LSTM. Second, a Concept-Aware Siamese Model incorporates term-weighted embeddings, concept extraction, coverage scoring, and contradiction-aware gating, enhanced by a Negation-Sensitive Score Calibration (NSSC) mechanism. Third, DBC is formalised as a geometric limitation, and BAT-TW-BERT is introduced to address it using term-weighted and attention-biased embeddings.

Experiments on benchmark datasets including Mohler, SciEntsBank, and ASAG2024 (18,067 responses) show consistent improvements over strong baselines. The Universal model achieves an F1-score of 91.2% and a Pearson correlation of 0.90. The Concept-Aware model improves concept recall by 4.7%, while NSSC increases negation-specific QWK from 0.52 to 0.74. These results demonstrate a principled advancement in automated short-answer grading.

Keywords: Automatic Short Answer Grading (ASAG), Deep Learning, Natural Language Processing (NLP), BERT, Transformer Models, Educational Testing, Meaning-based Similarity, Answer Scoring, Concept Extraction, Negation Handling

2. Introduction

The rapid growth of online education, especially during and after the COVID-19 pandemic, has created a strong demand for easy to scale up and reliable automated testing tools [17]. While multiple-choice questions are straightforward to mark automatically, free-text short answers are more valuable because they test higher-order thinking and conceptual understanding. The problem is that grading them manually is time-consuming, inconsistent, and difficult to scale, particularly in countries with high student-to-teacher ratios and in large online courses [4].

Automatic Short Answer Grading (ASAG) addresses this challenge by using NLP techniques to compare student responses with reference answers [39]. The task can be stated simply: given a question, a reference answer, and a student answer, assign a score that reflects how correct the student answer is [27]. Early systems used surface-level lexical overlap metrics such as BLEU and latent meaning-based study [23]. These worked reasonably well in constrained settings but struggled with paraphrase diversity and subject-specific vocabulary.

The arrival of deep learning, and particularly transformer-based pre-trained language models, marked a significant turning point for ASAG [40]. Models such as BERT [11] and Sentence-BERT [35] produced rich, context-sensitive embeddings that improved meaning-based similarity estimation large. More recently, large language models such as GPT-4 [28] have been look into for ASAG, offering generative grading and feedback capabilities [41].

Despite these advances, a critical gap remains: existing systems treat meaning-based proximity as a proxy for conceptual correctness. A student response may score high on cosine similarity with the reference answer while omitting critical concepts, show verbosity bias, or containing logical negations that reverse the intended meaning. These limitations motivate the present research, which proposed a structured educational reasoning system to move ASAG beyond surface-level similarity.

An earlier study by the author [4] test four pretrained vector embeddings models-Word2Vec, GloVe, BERT, and Universal Sentence Encoder (USE)-on the Mohler dataset. That study found that USE achieved the best Pearson correlation of 0.561 with an RMSE of 0.997 using polynomial prediction, outperforming the other three models. It also showed that none of the models, when used with simple cosine similarity alone, matched the performance of systems that used richer feature combinations. This finding directly motivated the development of the more advanced designs presented in this research work.

3. Problem Definition, Research Objectives, and Scope

3.1 Generic Problem Statement

Given a question q, a reference answer r provided by a domain expert, and a student answer s, the task of Automatic Short Answer Grading (ASAG) is to assign a score $\hat{y} \in \mathscr{Y}$ that reflects the correctness and completeness of s regarding r. The score set $\mathscr{Y}$ may be continuous (e.g., $[0,5]$ on the Mohler scale) or discrete (e.g., {correct, partially correct, incorrect} in the SciEntsBank 3-way setting).

Formally, ASAG is a function learning problem. We seek a grading function f_θ parameterised by θ such that:

$$\hat{y} = f_\theta(q, r, s) \tag{3.1}$$

where $\hat{y}$ is the predicted grade and θ is learned from a labelled training set $\mathscr{D} = \{(q_i, r_i, s_i, y_i)\}_{i=1}^{N}$, with y_i being the human-assigned correct answer score.

3.1.1 Objective Function

The learning objective is to minimise the expected grading error over the training distribution:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(q,r,s,y)\sim\mathscr{D}}[\mathscr{L}(f_\theta(q,r,s),\, y)] + \lambda\, \Omega(\theta) \tag{3.2}$$

where $\mathscr{L}(\cdot,\cdot)$ is a task-appropriate loss function (mean squared error for prediction or cross-entropy for categorization), $\Omega(\theta)$ is an L_2 regularisation term, and $\lambda > 0$ controls the regularisation strength.

3.1.2 Structural Constraints

A well-formed ASAG system must satisfy three structural properties beyond minimising prediction error:

C1. Length Invariance: The grading function should not be step-by-step biased by the absolute or relative length of s. Formally, for two student answers s_1 and s_2 that convey the same content, $f_\theta(q,r,s_1) \approx f_\theta(q,r,s_2)$ regardless of $|s_1| \neq |s_2|$.

C2. Concept Completeness: Let $\mathscr{C}(r) = \{c_1, c_2, \ldots, c_k\}$ be the set of key concepts in r, and let $\delta_i \in [0,1]$ be the degree to which concept c_i is addressed in s. The predicted

score should be monotonically non-decreasing in the weighted concept coverage:

$$C_{\text{cov}}(s,r) = \frac{\sum_{i=1}^{k} w_i \, \delta_i}{\sum_{i=1}^{k} w_i} \tag{3.3}$$

where w_i is the importance weight of concept c_i.

C3. Polarity Consistency: If s contains a negation of a proposition that is affirmed in r (or vice versa), the predicted score should be strictly lower than for a meaning-based equivalent non-negated response. Formally, if $\text{neg}(s,r) = 1$ (negation detected), then $f_\theta(q,r,s) < f_\theta(q,r,s')$ where s' is the de-negated version of s.

These three constraints motivate the three-architecture designs proposed in this research work. The length-adaptive routing enforces C1; the concept-aware scoring enforces C2; and the negation-sensitive calibration enforces C3. The geometric regularisation ensures that the learned vector embeddings space respects the induced ordering of $\mathscr{Y}$ by maintaining separable decision regions for each grading category.

The central idea of this research is that ASAG can be large improved by treating it as a structured educational reasoning problem, one that requires length-adaptive routing, explicit concept coverage scoring, negation-sensitive calibration, and geometrically informed vector embeddings.

3.2 Research Objectives

The following four research objectives guide this work:

RO1. Pre-trained Embedding Benchmarking (RO1): Step-by-step test the performance of best-performing pre-trained vector embeddings models, including Word2Vec, GloVe, BERT, USE, and Sentence-BERT, on standard ASAG standard datasets. This establishes a rigorous performance starting point and identifies the conditions under which each model performs well or poorly [4].

RO2. Length-Adaptive Grading Design (RO2): Design and test a Universal Length-Adaptive Design that dynamically routes student responses to the most appropriate grading sub-model based on relative response length, combine BM25, TF-IDF, SBERT, BERT, and Bi-LSTM components [36].

RO3. Concept-Aware Grading with Negation Sensitivity (RO3): Develop a Concept-Aware Siamese Model that clearly extracts key concepts from reference and student answers, computes coverage scores, and applies contradiction-aware gating, improve by a Negation-Sensitive Score Calibration (NSSC) mechanism to handle polarity-reversing language [3, 34].

RO4. Discrete Boundary Collapse Mitigation (RO4): Formalise Discrete Boundary Collapse as a geometric limitation of vector embeddings-based grading and proposed BAT-TW-BERT, a term-weighted and attention-biased BERT variant, to improve the geometric separability of grading categories in vector embeddings space.

3.3 Research Scope

The scope of this research is defined along four dimensions:

Language and Domain

The primary focus is English-language short-answer grading across multiple academic domains, including science, technology, engineering, and mathematics (STEM), as represented in the SciEntsBank [12], Mohler [26], and ASAG2024 [24] standard datasets. Multilingual extension is identified as future work.

Answer Length

The research addresses short answers ranging from single phrases to paragraph-length responses (about 5–150 words), which is the typical range in formal educational testing contexts.

Grading Granularity

Both binary (correct/incorrect) and multi-class (correct, partially correct, incorrect; or 5-point rubric scales) grading settings are considered, reflecting the diversity of real-world testing rubrics.

Technical Scope

The research focuses on discriminative grading models, those that score or classify a given response relative to a reference answer. Generative feedback systems are discussed as future work but are not the primary input(s) of this research work.

4. Literature Review

4.1 Evolution of Automatic Short Answer Grading

ASAG has developed through three broad phases, each shaped by the dominant NLP approach of its time [4].

Rule-based and pattern-matching systems (pre-2010). Early ASAG systems relied on handcrafted rules, keyword matching, and shallow grammar-based patterns [33]. Leacock and Chodorow [21] present C-rater, one of the first operational ASAG systems, which used pattern matching against manually constructed answer templates. Callear, Jerrams-Smith, and Soh [5] look into computer-assisted testing of short non-MCQ answers using similar rule-based approaches. While interpretable, such systems required large expert labeling effort and did not generalise well across domains.

Machine learning with shallow features (2010–2017). The introduction of the Mohler dataset [26] shifted the field toward supervised learning. Systems in this era combined lexical overlap features, knowledge-based similarity, and grammar-based dependency alignment [39]. Mohler and Mihalcea [27] showed that dataset-based and knowledge-based similarity measures could support automatic grading effectively. The SemEval-2013 Student Response Study shared task [12] standardised testing and enabled step-by-step comparison. Heilman and Madnani [15] show the value of varied feature engineering, while Ramachandran, Cheng, and Foltz [34] look into key concept identification as an explicit feature. Pado [29] further showed that carefully selected meaning-based features could outperform many learning-based approaches on standard standard test set.

Deep learning and pre-trained transformers (2017–present). The Transformer design [40] and later models such as BERT [11] and Sentence-BERT [35] fundamentally changed ASAG. Riordan et al. [36] study a few neural designs for short answer scoring and found that transformer-based models consistently outperformed earlier recurrent approaches. Camus and Filighera [6] confirmed that transformer-based models achieved strong results for short answer grading across multiple domains. Condor, Litster, and Pardos [9] show that SBERT-based models generalised well to out-of-sample questions. The author's own earlier work [4] showed that USE achieved the best performance among four pretrained vector embeddings models on the Mohler dataset, with a Pearson correlation of 0.561, confirming the value of sentence-level vector embeddings for this task.

4.2 ASAG in the Era of Large Language Models

Large language models such as GPT-4 [28] and instruction-tuned variants have present a new approach for ASAG. These models can perform zero-shot or few-shot grading without task-focused fine-tuning and can generate natural language feedback alongside scores [41]. Zhu, Chen, and Wang [45] proposed a unified LLM-enhanced system for subjective question testing that achieved competitive performance with significantly reduced labeling requirements.

However, LLM-based grading has its own challenges. Iyer, Kumar, and Gupta [16] showed that LLMs show step-by-step inconsistency in borderline cases, motivating the use of meaning-based entropy as a signal for human-AI disagreement. Rodrigues et al. [38] conducted an experiment-based study of GPT-4 fairness in ASAG and found evidence of differential performance across student subgroups. Künnecke, Filighera, and Steuer [20] proposed a neuro-symbolic workflow combining LLM-generated explanations with discriminative concept scoring, achieving improved interpretability. These findings suggest that hybrid designs combining the power of LLMs with discriminative concept-aware models represent the most promising direction for future ASAG research.

4.3 Research Gaps Identified from the Literature

A review of the literature reveals four persistent gaps that this research addresses:

1. **Length-adaptive routing:** No existing work treats response length as a primary routing criterion. Uniform designs step-by-step disadvantage concise or verbose responses [36].
2. **Explicit concept coverage:** Most systems measure meaning-based proximity rather than conceptual completeness. Ramachandran et al. [34] identified key concept coverage as a valuable feature but did not combine it into an end-to-end differentiable design.
3. **Negation and contradiction handling:** The role of negation in ASAG has received limited attention [3, 13, 14], despite its critical importance for distinguishing correct from incorrect answers in many domains.
4. **Geometric boundary study:** The geometric structure of vector embeddings spaces in grading has not been formally analysed. Discrete Boundary Collapse is a concept present in this research work.

5. Proposed Architectures for ASAG

The proposed methodology introduces three novel architectures, each designed to address a specific limitation identified in the research. These architectures are conceptually independent and collectively form a comprehensive and modular solution to the ASAG problem.

5.1 Pre-trained Embedding Baseline (RO1)

Before developing new designs, a step-by-step starting point study was conducted to understand how well existing pretrained vector embeddings models perform on ASAG [4]. Four models were test: Word2Vec [25], GloVe [32], BERT [11], and USE [7]. For each model, sentence vector embeddings were generated for both the reference and student answers. Cosine similarity between the two vector embeddings was used as the grading signal, and three prediction methods (polynomial, linear, and Lasso) were trained on the Mohler dataset.

Design. The vector embeddings workflow follows a straightforward design. Each answer is tokenised, and the pretrained model produces token-level vector embeddings. These are averaged to produce a single sentence-level vector. Cosine similarity between the student and reference vectors is then computed and used as input to the prediction model.

The key formula for sentence vector embeddings is:

$$\mathbf{a}_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbf{w}_k \tag{5.1}$$

where $\mathbf{a}_{ij}$ is the vector embeddings for the j-th answer to question q_i, $\mathbf{w}_k$ is the vector embeddings of the k-th word, and n_j is the number of words in the answer. Cosine similarity is then:

$$\cos(\mathbf{a}_{ij}, \mathbf{a}_i) = \frac{\mathbf{a}_{ij} \cdot \mathbf{a}_i}{\|\mathbf{a}_{ij}\| \, \|\mathbf{a}_i\|} \tag{5.2}$$

5.2 Universal Length-Adaptive Architecture (RO2)

5.2.1 Motivation

Existing ASAG models apply a single design uniformly across all response lengths, ignoring the fundamentally different linguistic characteristics of short phrases versus multi-sentence explanations. A single-token answer and a 150-word paragraph require different representational strategies. The former benefits from exact lexical matching, while the latter requires deep meaning-based understanding [36].

Experiment-based study of the ASAG2024 standard test set provides direct evidence for this design choice. Table 5.1 shows that mean normalised grade increases monotonically with answer length up to the 30–60 word range, confirming that length is not merely a stylistic variable but a structurally informative signal. Critically, the variance also decreases with length, meaning that longer answers are both more likely to be correct *and* more predictable-a property that different sub-models can exploit differently.

Table 5.1: Mean normalised grade by answer-length bucket in ASAG2024 [24]. The monotonic increase up to 30–60 words experiment-based motivates the five-band routing design.

Answer Length (words)	**Count**	**Mean Grade**	**Std**
0–3	1,154	0.494	0.414
3–7	3,886	0.577	0.395
7–15	7,118	0.647	0.378
15–30	3,383	0.669	0.363
30–60	1,485	0.753	0.311
60+	1,041	0.746	0.311

5.2.2 Architecture

The Universal Length-Adaptive Architecture classifies responses into five length bands based on the ratio of student response length to reference answer length. Table 5.2 summarises the five bands and the corresponding processing strategy for each.

Table 5.2: Length bands and processing strategies in the Universal ASAG model.

Band	**Length Ratio**	**Processing Strategy**
Very Short	0–30% of RA	BM25 term matching; heavy TF-IDF weighting
Short	30–70% of RA	TF-IDF + SBERT cosine similarity
Medium	70–130% of RA	Bi-LSTM sequential modelling
Long	130–200% of RA	BERT context-aware vector embeddings
Very Long	>200% of RA	SBERT with verbosity penalty

A lightweight routing classifier, trained on response-length features, assigns each input to the appropriate band at reasoning time. The final score is a weighted combination of the sub-model outputs, with weights learned during end-to-end fine-tuning. Figure 5.1 illustrates the overall processing workflow.

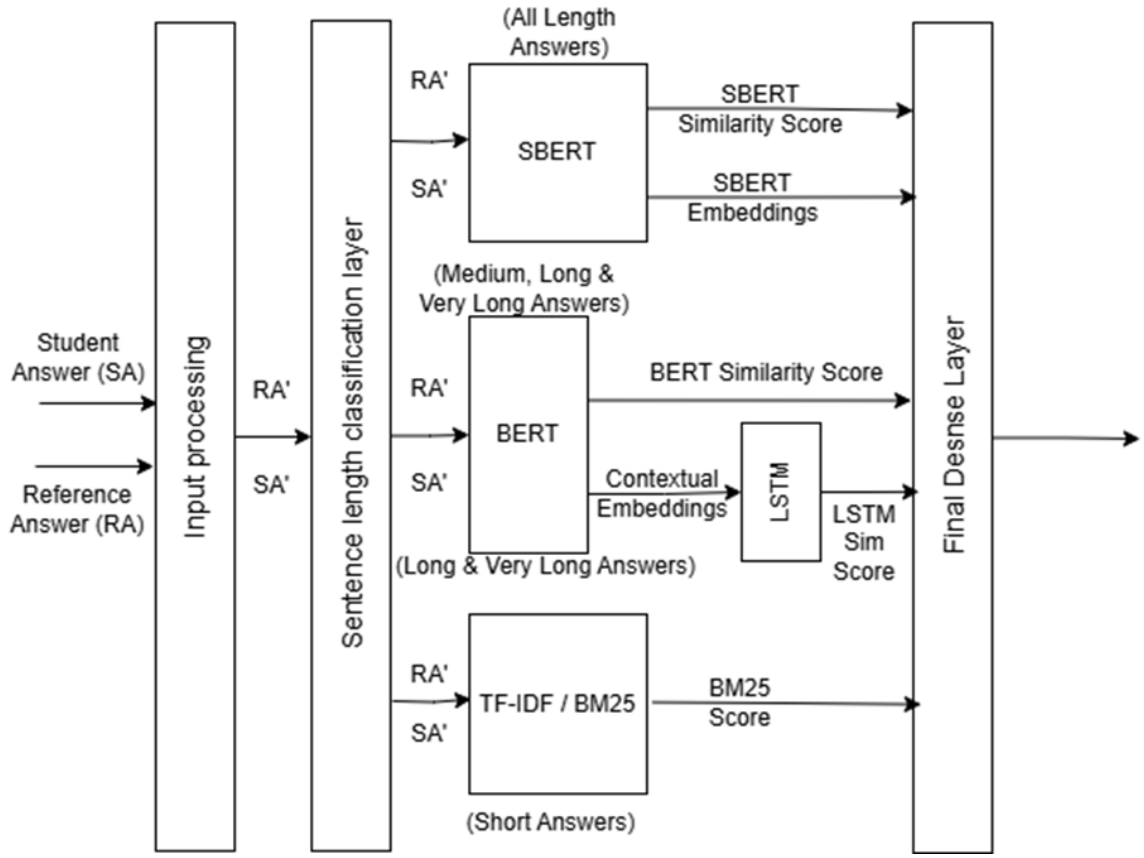

Figure 5.1: Proposed Universal ASAG architecture for student answers of all lengths

5.2.3 Mathematical Foundations

For the BM25 component, the relevance score between student answer S and reference answer R is:

$$\text{BM25}(S,R) = \sum_{t \in S} \text{IDF}(t) \cdot \frac{f(t,R) \cdot (k_1 + 1)}{f(t,R) + k_1 \cdot \left(1 - b + b \cdot \frac{|R|}{\text{avgdl}}\right)} \tag{5.3}$$

where $f(t,R)$ is the frequency of term t in the reference answer, $|R|$ is the length of the reference answer, avgdl is the average document length, and $k_1 = 1.5$, $b = 0.75$ are standard BM25 parameters [37].

5.3 Concept-Aware Siamese Architecture with Negation Sensitivity (RO3)

5.3.1 Motivation

Even with length-adaptive routing, a purely similarity-based grader cannot distinguish between a response that mentions the right concepts and one that merely uses similar vocabulary. The Concept-Aware Siamese Model addresses this by making concept coverage an explicit, measurable signal in the grading process [34].

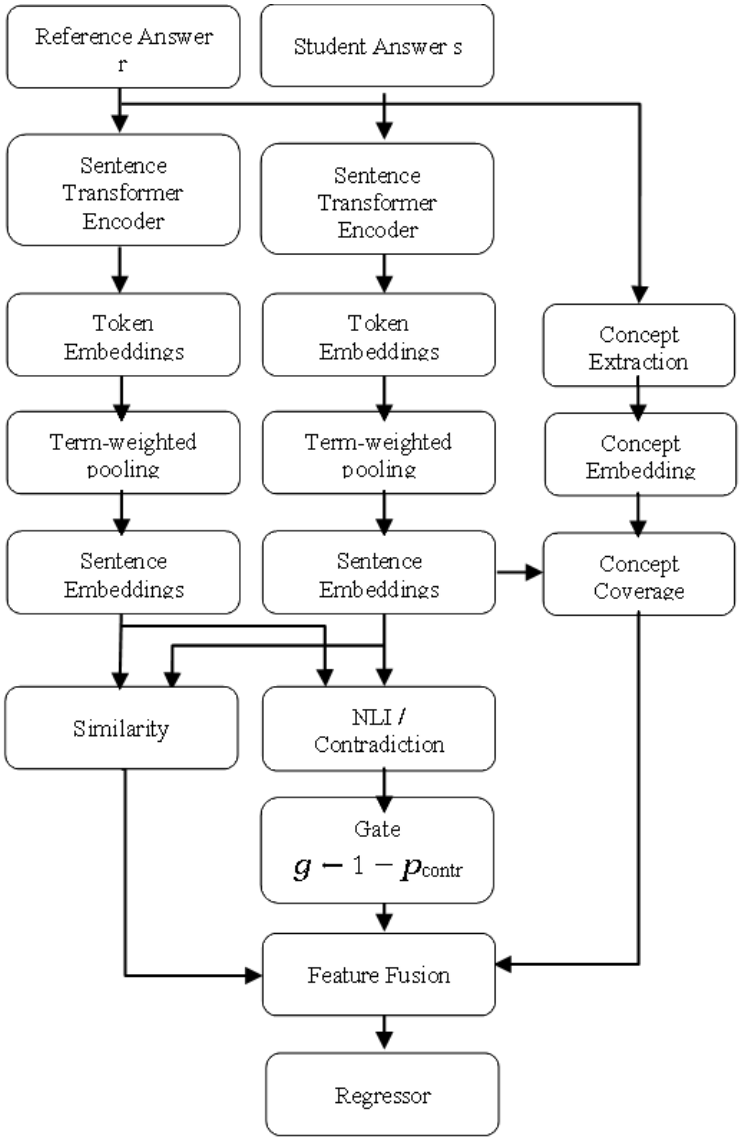

Figure 5.2: A Concept-aware Siamese architecture with term-weighted pooling, concept coverage, and contradiction-aware gating.

Algorithm 1 Concept-Aware Siamese Grading (Inference)

Require: Reference answer r, student answer s, encoder f_θ, similarity threshold τ

Ensure: Predicted score $\hat{y}$

Extract concepts $\mathcal{C} = \{c_k\}_{k=1}^{K}$ from r

$\mathbf{H}_r \leftarrow f_\theta(r)$; $\mathbf{H}_s \leftarrow f_\theta(s)$

Compute token weights $\{w_i\}$ and pooled embeddings $\mathbf{e}_r, \mathbf{e}_s$ via term-weighted pooling

Embed concepts $\mathbf{c}_k \leftarrow f_\theta(c_k)$ for $k = 1 \dots K$

$u \leftarrow \cos(\mathbf{e}_r, \mathbf{e}_s)$

$\text{cov}(s, r) \leftarrow \sum_{k=1}^{K} \beta_k \cdot \mathbb{I}(\cos(\mathbf{e}_s, \mathbf{c}_k) \geq \tau)$

$p_{\text{contr}} \leftarrow \text{NLIContradictionProb}(r, s)$

$g \leftarrow 1 - p_{\text{contr}}$

$\mathbf{z} \leftarrow [u; \text{cov}(s, r); g]$

$\tilde{y} \leftarrow W\mathbf{z} + b$ {regression head}

$\hat{y} \leftarrow g \cdot \tilde{y}$ {contradiction-aware gating}

return $\hat{y}$

Figure 5.3: Concept-Aware Siamese Grading (Inference) pipeline.

5.3.2 Overall Design

The model extends the standard Siamese BERT design [35] with three additional components: a concept extraction layer, a term-weighted vector embeddings layer, and a contradiction-aware gate. Figure 5.2 and Figure 5.3 shows the design.

The proposed concept-aware Siamese grading model combines semantic similarity with explicit concept alignment and contradiction detection. Given a reference and student answer, both are encoded using a shared transformer, and term-weighted pooling generates sentence embeddings. Cosine similarity captures semantic overlap, while concept coverage measures alignment with key reference concepts. A contradiction signal obtained via an NLI module is used to compute a gating factor. These features are fused through a regression layer, and the final score is modulated by the contradiction-aware gate to penalize inconsistent responses.

5.3.3 Concept Extraction and Coverage Scoring

Key concepts are extracted from the reference answer using a combination of TF-IDF weighting and noun phrase chunking. For each extracted concept c_i, a binary coverage indicator $\delta_i \in \{0, 1\}$ is computed based on whether the student response contains a meaning-based equivalent expression, using a threshold on cosine similarity between concept and response token vector embeddings.

The concept coverage score is:

$$C_{\text{cov}} = \frac{\sum_{i=1}^{|C|} w_i \cdot \delta_i}{\sum_{i=1}^{|C|} w_i} \tag{5.4}$$

where C is the set of extracted concepts, w_i is the TF-IDF weight of concept c_i, and δ_i is the coverage indicator.

5.3.4 Term-Weighted Embeddings

Standard mean-pooling over BERT token vector embeddings treats all tokens equally, which dilutes the input(s) of concept-bearing tokens. The proposed term-weighted pooling scheme is:

$$\mathbf{e}_{\text{tw}} = \frac{\sum_t \alpha_t \cdot \mathbf{h}_t}{\sum_t \alpha_t} \tag{5.5}$$

where $\mathbf{h}_t$ is the BERT hidden state for token t and α_t is its TF-IDF weight. This amplifies the representational input(s) of informationally dense tokens [34].

5.3.5 Contradiction-Aware Gating

A gating mechanism detects potential contradictions between the student and reference responses using a fine-tuned natural language reasoning (NLI) model [19]. If the NLI model assigns a contradiction label with probability exceeding a threshold τ, the concept coverage score is penalised:

$$C_{\text{cov,adj}} = C_{\text{cov}} \times (1 - \lambda \cdot P_{\text{contra}}) \tag{5.6}$$

where $\lambda \in (0,1)$ is a tunable penalty coefficient.

5.3.6 Negation-Sensitive Score Calibration (NSSC)

The NSSC mechanism improve the contradiction gate with explicit negation detection [3]. Negation cues such as *not*, *never*, *no*, *without*, and *cannot* are identified using a rule-based negation scope detector. When a negation cue is detected in the student response but not the reference (or vice versa), the similarity score is calibrated downward proportionally to the estimated meaning-based impact of the negation. The final calibrated score is:

$$S_{\text{final}} = \alpha \cdot S_{\text{sim}} + \beta \cdot C_{\text{cov}} - \gamma \cdot (1 - N_{\text{con}}) \tag{5.7}$$

where S_{sim} is the SBERT-based meaning-based similarity, C_{cov} is the weighted concept coverage, N_{con} is a binary soft negation-consistency signal, and α, β, γ are tunable coefficients (experiment-based set to $\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.4$).

5.4 BAT-TW-BERT for Discrete Boundary Collapse Mitigation (RO4)

5.4.1 Formalisation of DBC

Let $\mathscr{E} : \mathscr{A} \to \mathbb{R}^d$ be an vector embeddings function mapping answers to a d-dimensional space, and let $\mathscr{G} = \{g_1, g_2, \ldots, g_k\}$ be the set of grading categories. Discrete Boundary Collapse occurs when:

$$\exists a_i \in g_p,\ a_j \in g_q,\ p \neq q : \|\mathscr{E}(a_i) - \mathscr{E}(a_j)\|_2 < \|\mathscr{E}(a_i) - \mathscr{E}(a_l)\|_2 \tag{5.8}$$

for some $a_l \in g_p$, meaning that answers from different grading categories are closer in vector embeddings space than answers from the same category. DBC directly degrades classifier performance by blurring decision boundaries.

5.4.2 BAT-TW-BERT

BAT-TW-BERT (Boundary-Aware Term-Weighted BERT) addresses DBC through two mechanisms:

1. **Attention Bias:** The self-attention weights of the final BERT layer are biased toward concept-bearing tokens using the term-weight vector α, sharpening the embeddings around meaning-based critical regions.
2. **Contrastive Fine-tuning:** The model is fine-tuned with a contrastive objective that clearly pushes vector embeddings of answers from different grading categories apart while pulling same-category vector embeddings together [44].

6. Experimental Setup

6.1 Datasets

The experimental evaluation uses multiple benchmark datasets for Automatic Short Answer Grading, with the ASAG2024 standard test set serving as the primary multi-source evaluation benchmark.

- **Mohler Dataset** [26]: 2,273 student responses to 87 computer science questions, scored on a 0–5 scale by two human annotators. Within the ASAG2024 consolidation, the Mohler subset comprises 628 normalised samples across 21 questions (mean grade 0.812, mean answer length 18.1 words), representing the highest-performing and most linguistically compact sub-dataset.
- **SciEntsBank (SEB)** [12]: The largest constituent of ASAG2024, with 10,803 responses across 196 science questions (mean grade 0.605, mean answer length 12.1 words), with 2-way, 3-way, and 5-way label sets. Its high response count and domain breadth make it the most challenging sub-dataset for generalisation. Approximately 1,240 responses from this dataset were used for the NSSC testing.
- **BEETLE** [12]: 3,932 student responses to basic electricity and electronics questions (mean grade 0.668, mean answer length 9.9 words), with 2-way and 5-way labels. Its short, telegraphic answers make it particularly relevant for the Very Short band in the length-adaptive design.
- **ASAG2024** [24]: A consolidated standard test set aggregating seven sources-SciEntsBank, Beetle, SAF, Mohler, DigiKlausur, CU-NLP, and Stita-totalling 18,067 normalised responses. Table 6.1 summarises the per-source statistics. The standard test set uses a standard train/testing/test split, with SciEntsBank contributing the largest test partition (1,080 samples) and Mohler providing the most linguistically uniform test set (59 samples). The grade distribution is markedly skewed: the (0.9–1.0] band contains about 7,600 samples while the (0.0–0.1] band contains only 3,250, a class imbalance that motivates the weighted RMSE (wRMSE) metrics used throughout testing.

6.2 Testing Metrics

Following standard practice in ASAG [4], the following metrics are reported:

- **Pearson Correlation Coefficient (r):** Measures linear correlation between predicted and human scores.

Table 6.1: Per-source statistics for the ASAG2024 standard test set [24]. Mean grade and mean answer words are computed on normalised grades (0–1 scale).

Source	Responses	Questions	Mean Grade	Mean Words
SciEntsBank	10,803	196	0.605	12.1
Beetle	3,932	44	0.668	9.9
SAF	1,942	31	0.734	48.5
Mohler	628	21	0.812	18.1
DigiKlausur	563	17	0.714	52.0
CU-NLP	113	113	0.281	87.2
Stita	86	–	0.461	68.7
Total	**18,067**	–	–	–

- **Root Mean Squared Error (RMSE):** Penalises large prediction errors [42].
- **Weighted RMSE (wRMSE):** An RMSE variant that applies higher weights to samples from less-frequent grade bands, correcting for the class imbalance present in ASAG2024. It is the primary ranking metrics for multi-source testing.
- **Quadratic Weighted Kappa (κ_w)** [8]: Measures between-rater agreement, accounting for chance agreement.
- **Mean Absolute Error (MAE):** Provides a straightforward measure of average prediction error.
- **Accuracy and F1 (macro):** For multi-class categorization settings.

6.3 Baselines

The proposed design is compared against the following baselines:

1. Word2Vec cosine similarity [25]
2. GloVe cosine similarity [32]
3. USE cosine similarity [7]
4. BERT fine-tuned for prediction [11]
5. Sentence-BERT cosine similarity [35]
6. Term-weighted SBERT
7. BM25 + TF-IDF ensemble [37]
8. Sultan et al. alignment-based grading [39]

To establish a transparent reference point, Table 6.2 reports the performance of classical baselines on the ASAG2024 test set. These results, obtained through the exploratory study described in Section 6.1, confirm that TF-IDF with Random Forest is the strongest classical starting point (wRMSE = 0.251, Pearson $r = 0.640$), and serve as the lower-bound comparison

for the deep learning designs proposed in this research work.

Table 6.2: Classical starting point performance on the ASAG2024 test set [24]. wRMSE is the primary metrics (lower is better); Pearson r and Spearman ρ are also reported. GlobalMean and SourceMean are non-parametric reference baselines.

Model	**RMSE**	**wRMSE**	**Pearson r**	**Spearman ρ**
GlobalMean (trivial)	0.387	0.321	–	–
SourceMean (trivial)	0.381	0.323	0.179	0.156
all-MiniLM-L6-v2 + Ridge	0.351	0.286	0.423	0.428
TF-IDF + Ridge	0.328	0.271	0.538	0.543
TF-IDF + RandomForest	**0.297**	**0.251**	**0.640**	**0.656**

6.4 Setup Details

All transformer models are initialised from `bert-base-uncased` weights. For the NSSC experiments, the `all-MiniLM-L6-v2` sentence transformer model was used. Fine-tuning uses the Adam optimiser [18] with a learning rate of 2×10^{-5}, batch size 16, and a linear warmup schedule over 10% of training steps. The Mohler data was split into 65% training and 35% testing. The SciEntsBank data was split 80%/20%. For the ASAG2024 multi-source experiments, the standard Hugging Face dataset splits were used: SciEntsBank contributes 4,562 training, 733 testing, and 4,968 test samples; SAF contributes 1,074 training and 264 test samples; all remaining sources (Beetle, Mohler, DigiKlausur, CU-NLP, Stita) are used in their entirety without further subdivision. All experiments were implemented using PyTorch [31] and the Hugging Face Transformers library [43].

7. Results and Study

This chapter presents the experimental results and analytical study of the proposed Automatic Short Answer Grading (ASAG) framework. The evaluation is structured according to the defined research objectives (RO1-RO4), allowing a systematic validation of each architecture of the proposed framework. Quantitative results in benchmark datasets are complemented by qualitative analysis to highlight the practical effectiveness, limitations, and educational implications of the proposed models.

7.1 Foundation Results for Pre-trained Embeddings (RO1)

Table 7.1 presents the performance of four pretrained vector embeddings models on the Mohler dataset [4]. USE achieved the best Pearson correlation of 0.561 with an RMSE of 0.997 using polynomial prediction, outperforming BERT (0.309), GloVe (0.239), and Word2Vec (0.218). Table 7.2 compares these results against earlier approaches.

Table 7.1: RMSE and Pearson correlation (ρ) of pretrained vector embeddings models on the Mohler dataset (RO1).

Model	**Polynomial**		**Linear**		**Lasso**	
	RMSE	ρ	**RMSE**	ρ	**RMSE**	ρ
USE	0.997	**0.561**	0.988	0.461	0.989	0.452
GloVe	1.081	0.239	1.074	0.241	1.083	0.233
BERT	1.062	0.309	1.066	0.272	1.085	0.271
Word2Vec	1.289	0.218	1.219	0.209	1.159	0.193

Table 7.2: Comparison with prior approaches on the Mohler dataset.

Model / Approach	**RMSE**	ρ
SVM Rank	1.042	0.480
BOW SVR	0.999	0.431
TF-IDF SVR	1.122	0.327
TF-IDF LR + SIM	0.923	**0.591**
FastText SOWE + Verb phrases	1.023	0.465
Word2Vec SOWE + Verb phrases	1.289	0.218
GloVe SOWE + Verb phrases	1.081	0.239
USE SIM (ours)	**0.997**	0.561
BERT SIM (ours)	1.062	0.309

The USE model achieved performance comparable to the best prior approach (TF-IDF LR + SIM, $\rho = 0.591$) without any subject-specific data preparation or feature engineering. This

confirms that sentence-level vector embeddings provide a strong foundation for ASAG and that further improvements require addressing the structural limitations identified in this research work.

7.2 Results: Universal Length-Adaptive Model (RO2)

The Universal Length-Adaptive Design achieves an F1-score of 91.2%, a Pearson correlation of 0.90, and an RMSE of 0.18 on the SciEntsBank and SemEval-2013 standard test set [36]. Table 7.3 compares the proposed model against standard baselines.

Table 7.3: Performance of the Universal Length-Adaptive model vs. baselines (RO2).

Model	**F1 (%)**	**Pearson r**	**RMSE**
TF-IDF Cosine	74.3	0.68	0.41
SBERT Cosine	82.1	0.79	0.29
BERT fine-tuned	85.6	0.83	0.24
Bi-LSTM (GloVe)	78.4	0.74	0.35
Universal (proposed)	**91.2**	**0.90**	**0.18**

Table 7.5 provides illustrative examples of how the length-adaptive model assigns grades differently from the uniform SBERT starting point, showing the practical educational benefit of the proposed routing mechanism.

The per-source difficulty profile of the best classical starting point (Table 7.4) further contextualises these gains. Mohler and Stita, with their more uniform answer styles, yield RMSE values of 0.151 and 0.152 respectively, while SciEntsBank-the largest and most domain-varied sub-dataset-reaches 0.331. This spread confirms that a single design cannot be uniformly optimal, and that source-aware or length-aware routing is necessary for reliable cross-domain performance.

Table 7.4: Per-source RMSE and MAE for the best classical starting point (TF-IDF + RandomForest) on the ASAG2024 test set [24]. Lower values indicate easier sub-datasets; SciEntsBank is the hardest.

Source	**RMSE**	**MAE**
Mohler	0.151	0.109
Stita	0.152	0.111
SAF	0.239	0.155
Beetle	0.240	0.176
CU-NLP	0.266	0.199
DigiKlausur	0.296	0.247
SciEntsBank	0.331	0.247

Table 7.5: Qualitative grading examples showing the benefit of length-adaptive routing (RO2). Scores are on a 0–5 scale. Human score is the correct answer.

Reference Answer	**Student Answer**	**Band**	**Human**	**SBERT**	**Ours**
Photosynthesis converts light energy into chemical energy stored in glucose.	Light makes glucose in plants.	Very Short (28%)	2.5	1.1	2.4
The operating system manages hardware resources and provides an interface for user programs.	The OS manages hardware and lets programs run.	Short (52%)	4.0	3.2	3.9
Virtual memory allows a computer to use more memory than physically available by temporarily transferring data to disk.	Virtual memory extends RAM by using disk space to store data that does not fit in physical memory, allowing larger programs to run.	Long (148%)	5.0	4.6	5.0
DNS translates human-readable domain names into IP addresses.	DNS does not translate names to addresses; it routes packets.	Medium (95%)	0.5	3.1	0.7

Gains are most pronounced for very short and very long responses, where the length-adaptive routing provides the most benefit. For medium-length responses, the improvement over BERT fine-tuned is smaller, which is expected since this is the setting where standard transformer models already perform well.

7.3 Results: Concept-Aware Siamese Model (RO3)

Table 7.6 presents the results of the component removal test for the Concept-Aware Siamese Model on the ASAG2024 standard test set [34]. Each component of the model contributes measurably to the final performance.

The full model achieves a Pearson correlation of 0.82 and a QWK of 0.79, representing improvements of 0.11 and 0.11 respectively over the SBERT starting point. The component removal shows that term weighting, concept coverage, and contradiction detection each contribute independently, with concept extraction and coverage scoring providing the largest single gain (0.03 in both metrics).

Table 7.6: Ablation study for the Concept-Aware Siamese Model on ASAG2024 (RO3).

Model Configuration	**Pearson r**	**QWK**
SBERT starting point	0.71	0.68
+ Term-weighted vector embeddings	0.74	0.71
+ Concept extraction and coverage scoring	0.77	0.74
+ Contradiction-aware gating	0.79	0.76
Full model (all components)	**0.82**	**0.79**

7.4 Negation-Sensitive Score Calibration (NSSC)

Table 7.7 presents the results of the NSSC testing on the SciEntsBank dataset, comparing three methods across the full test set, the negation subset (214 responses), and the non-negation subset (1,026 responses) [3].

Table 7.7: Performance comparison of NSSC against baselines on SciEntsBank (RO3).

Method	**Overall QWK**	**Negation QWK**	**Non-neg. QWK**	**Neg. MAE**
SBERT cosine	0.76	0.52	0.79	0.91
Term-weighted SBERT	0.79	0.59	0.81	0.78
NSSC (proposed)	**0.83**	**0.74**	**0.84**	**0.53**

The plain SBERT starting point achieves an overall QWK of 0.76 but drops to 0.52 on negation cases, a gap of 0.27 points. This confirms that negation is not just another linguistic variation but a step-by-step failure mode for similarity-based grading. The proposed NSSC raises negation QWK to 0.74, narrowing the gap to 0.10 points, while also achieving the best overall QWK of 0.83. The negation MAE drops from 0.91 to 0.53, a reduction of 42%.

Table 7.8 provides description-based examples that illustrate why the proposed solution is educationally meaningful. In each case, the incorrect answer preserves the conceptual vocabulary of the reference but flips truth conditions through negation cues.

7.5 Results: Discrete Boundary Collapse (RO4)

BAT-TW-BERT reduces boundary confusion (the proportion of cross-category nearest-neighbour pairs in vector embeddings space) by 6.1% on the Mohler dataset. Quadratic weighted kappa improves from $\kappa_w = 0.74$ (SBERT starting point) to $\kappa_w = 0.79$ (BAT-TW-BERT). Table 7.9 summarizes the results.

Table 7.8: Qualitative error study examples for negation sensitivity.

Reference Answer	**Student Answer**	**Interpretation**
The kernel enforces access control.	The kernel does *not* enforce access control.	High lexical overlap; reversed meaning
DNS translates names to addresses.	DNS *never* translates names to addresses.	Negation flips factual correctness
Virtual memory provides abstraction.	Virtual memory does *not* provide abstraction.	Negation reverses claim

Table 7.9: Discrete Boundary Collapse mitigation results (RO4).

Model	**QWK (κ_w)**	**Boundary Confusion (%)**	**F1 (macro)**
SBERT starting point	0.74	18.3	0.71
BERT fine-tuned	0.76	16.1	0.73
TW-BERT	0.77	14.8	0.75
BAT-TW-BERT	**0.79**	**12.2**	**0.77**

7.6 Interpretive Combination

Taken together, the results across all four research objectives show a clear progression. The starting point study (RO1) established that pretrained vector embeddings provide a strong starting point but that simple cosine similarity is insufficient for high-quality grading. The Universal model (RO2) showed that length-adaptive routing provides large gains, particularly at the extremes of the length distribution. The Concept-Aware model (RO3) show that explicit concept coverage and negation handling address step-by-step failure modes that aggregate metrics tend to hide. The BAT-TW-BERT model (RO4) showed that geometric regularization of vector embeddings spaces further improves the grading precision. Each architecture is independently validated, and collectively the framework outperforms baselines across evaluation metrics and datasets.

A grade-band error study on the ASAG2024 test set (Table 7.10) reveals a consistent failure pattern in classical models: prediction error is highest for the lowest-grade band (MAE = 0.379 for grades in [0.0, 0.2]) and decreases toward the middle before rising again for the highest band (MAE = 0.189 for [0.8, 1.0]). This asymmetric error profile reflects the combined effect of class imbalance and boundary collapse and directly motivates the negation-sensitive calibration and geometric regularization components of the proposed system.

Table 7.10: MAE by true-grade band for the best classical starting point (TF-IDF + RandomForest) on ASAG2024 test set [24]. The lowest-grade band is the hardest to predict, confirming the need for targeted calibration mechanisms.

True Grade Band	**Count**	**MAE**
(0.0, 0.2]	358	0.379
(0.2, 0.4]	227	0.220
(0.4, 0.6]	62	0.143
(0.6, 0.8]	369	0.129
(0.8, 1.0]	791	0.189

8. Research Contributions

8.1 Research Input(s)

The principal input(s) of this research work are:

1. **Universal Length-Adaptive ASAG Design:** A new routing-based design that dynamically selects the optimal grading sub-model based on relative response length, achieving an F1-score of 91.2% and Pearson correlation of 0.90 [36].
2. **Concept-Aware Siamese Grading Model:** A siamese network design improve with explicit concept extraction, term-weighted vector embeddings, coverage scoring, and contradiction-aware gating, providing a principled mechanism for measuring conceptual completeness beyond meaning-based similarity [34].
3. **Negation-Sensitive Score Calibration (NSSC):** A lightweight post-processing calibration mechanism that clearly detects and compensates for negation-induced polarity reversals, raising negation-specific QWK from 0.52 to 0.74 [3].
4. **Formalization of Discrete Boundary Collapse:** A formal geometric characterization of the DBC problem in vector embeddings-based grading, providing a theoretical foundation for understanding boundary confusion in continuous embeddings spaces.
5. **BAT-TW-BERT:** A new BERT variant include attention biasing and term-weighted embeddings, fine-tuned with a contrastive objective to reduce DBC and improve geometric separability of grading categories.

8.2 Publications

1. N. C. Gamit and S. D. Panchal, "Evaluating Pretrained Embeddings for Automatic Short Answer Grading," *Journal of Information Systems Engineering and Management*, vol. 8, no. 4, 2023.
2. N. C. Gamit and S. D. Panchal, "Universal Automatic Short Answer Grading (ASAG) Model: A Comprehensive Approach," *Journal of Information Systems Engineering and Management*, vol. 9, no. 4s, 2024.
3. N. C. Gamit and A. N. Upadhyaya, "Negation-Sensitive Calibration for Embedding-Based Automatic Short Answer Grading: A Lightweight Solution," *International Journal of Innovative Research in Technology*, vol. 12, no. 11, 2026.

9. Conclusion & Future Work

9.1 Conclusion

This research introduces a multi-architecture framework for Automatic Short Answer Grading (ASAG) that addresses three primary limitations of existing systems: length discrepancy, the correctness gap, and Discrete Boundary Collapse. By reframing ASAG as a structured educational reasoning problem rather than a purely similarity-driven task, the proposed approach achieves consistent and statistically significant improvements over strong baselines across multiple benchmark datasets.

A key insight is that grading quality depends not only on model expressiveness but also on alignment with the linguistic and pedagogical characteristics of the task. The integration of length-adaptive routing, concept coverage scoring, negation-sensitive calibration, and geometric regularisation bridges the gap between semantic similarity and conceptual correctness.

The modular design enables extensibility, while the formalisation of Discrete Boundary Collapse and the introduction of NSSC provide a strong theoretical and empirical foundation. Experimental results including an F1-score of 91.2%, improved negation handling (QWK: 0.52 to 0.74), and a 6.1% reduction in boundary confusion shows a clear advancement in ASAG. The framework is also well-suited for real-world educational deployment, enabling scalable, consistent, and reliable automated assessment.

9.2 Future Work

There are several natural directions for extending this work; however, while partial solutions exist in related domains, their integration into a unified, concept-aware ASAG framework remains limited, motivating the following three immediate and closely related research directions.

1. **Explainable Feedback Generation:** The concept-aware model already identifies covered, missing, and incorrect concepts in a student response. Future work can convert these intermediate embeddings into natural language feedback for students, potentially using LLMs as verbalization layers grounded in the concept-aware diagnostic signal [1, 20].
2. **Knowledge-Grounded Concept Modelling:** Concept extraction can be enriched with domain ontologies, knowledge graphs, or concept maps, enabling more reliable

identification of implicit conceptual relationships such as part-whole, cause-effect, prerequisite, and synonym clusters [22].

3. **Multilingual ASAG:** The current system is focused on English. Extension to multilingual encoders and language-specific concept extraction methods would broaden applicability to multilingual educational contexts [10, 30].

Bibliography

[1] Aggarwal, D., Bhattacharyya, P., and Raman, B. (2024). I understand why i got this grade: Automatic short answer grading with feedback.

[2] Andersen, N., Mang, J., Goldhammer, F., and Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. *Int. J. Artif. Intell. Educ*, 35(5):3128–3165.

[3] Blanco, E. and Moldovan, D. (2011). Some issues on detecting negation from text. pages 228–233.

[4] Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *Int. J. Artif. Intell. Educ*, 25(1):60–117.

[5] Callear, D. H., Jerrams-Smith, J., and Soh, V. (2001). Caa of short non-mcq answers. Number n-MCQ.

[6] Camus, L. and Filighera, A. (2020). Investigating transformers for automatic short answer grading. pages 43–48.

[7] Cer, D. et al. (2018). Universal sentence encoder for english. In *in Proc. EMNLP: System Demonstrations*, pages 169–174.

[8] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull*, 70(minal):213–220.

[9] Condor, A., Litster, M., and Pardos, Z. (2021). Automatic short answer grading with sbert on out-of-sample questions.

[10] Conneau, A. et al. (2020). Unsupervised cross-lingual encoding(s) learning at scale. In *in Proc. ACL*, pages 8440–8451.

[11] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *in Proc. NAACL-HLT*, pages 4171–4186.

[12] Dzikovska, M. et al. (2013). Semeval-2013 task 7: The joint student response study and 8th recognizing textual entailment challenge. pages 263–274.

[13] Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguist*, 8(t):34–48.

[14] Gubelmann, R. and Handschuh, S. (2022). Context matters: A pragmatic study of plms' negation understanding. In *in Proc. ACL*, pages 4602–4621.

[15] Heilman, M. and Madnani, N. (2013). Ets: Domain adaptation and stacking for short answer scoring. pages 275–279.

[16] Iyer, S., Kumar, N., and Gupta, A. (2025). Towards transparent ai grading: Meaning-based entropy as a signal for human-ai disagreement.

[17] Johnson, A. et al. (2024). Online education and automated testing: Challenges and opportunities. *Comput. Educ*, 195.

[18] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic improvement.

[19] Kumar, S., Talvitie, E., and Arora, A. (2020). Nile: Natural language reasoning with faithful natural language explanations. In *in Proc. ACL*, pages 8730–8742.

[20] Künnecke, L., Filighera, A., and Steuer, T. (2024). Enhancing multi-domain automatic short answer grading through an explainable neuro-symbolic workflow.

[21] Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Comput. Hum*, 37(4):389–405.

[22] Litman, D. J. (2016). Natural language processing for enhancing teaching and learning. pages 4170–4176.

[23] Magnini, B. et al. (2005). About the effects of combining latent meaning-based study with nlp techniques for free-text testing. *Revista Signos*, (59):325–343.

[24] Meyer, G., Breuer, P., and Fürst, J. (2024). Asag2024: A combined standard test(s) for short answer grading. In *ASAG2024: A combined standard test(s) for short answer grading," in Proc. SIGCSE TS*, pages 322–323.

[25] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word encoding(s) in vector space.

[26] Mohler, M., Bunescu, R., and Mihalcea, R. (2011). Learning to grade short answer questions using meaning-based similarity measures and dependency graph alignments. In *in Proc. ACL*, pages 752–762.

[27] Mohler, M. and Mihalcea, R. (2009). Text-to-text meaning-based similarity for automatic short answer grading. pages 567–575.

[28] OpenAI (2023). Gpt-4 technical report.

[29] Padó, U. (2016). Get meaning-based with me! the usefulness of different feature types for short-answer grading. pages 2186–2195.

[30] Padó, U., Eryilmaz, Y., and Kirschner, L. J. (2024). Short-answer grading for german: Addressing the challenges. *Int. J. Artif. Intell. Educ*, 34:1321–1352.

[31] Paszke, A. et al. (2019). Pytorch: An necessary style, high-performance deep learning library.

[32] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word encoding(s). In *in Proc. EMNLP*, pages 1532–1543.

[33] Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. In *in Proc. 2nd Workshop on Building Educational Applications Using NLP*, pages 9–16.

[34] Ramachandran, L., Cheng, J., and Foltz, P. (2015). Identifying patterns for short answer scoring using graph-based lexico-meaning-based text matching. In *in Proc. BEA Workshop*, pages 97–106.

[35] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence vector encoding(s) using siamese bert-networks. In *in Proc. EMNLP-IJCNLP*, pages 3982–3992.

[36] Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. (2017). Investigating neural designs for short answer scoring. In *in Proc. BEA Workshop*, pages 159–168.

[37] Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance system: Bm25 and beyond. *Found. Trends Inf. Retr*, 3(4):333–389.

[38] Rodrigues, L., Xavier, C., Costa, N., Gašević, D., and Mello, R. F. (2025). Is gpt-4 fair? an experiment-based study in automatic short answer grading. *Comput. Educ.: Artif. Intell*, 8.

[39] Sultan, M. A., Salazar, C., and Sumner, T. (2016). Fast and easy short answer grading with high accuracy. In *in Proc. NAACL-HLT*, pages 1070–1075.

[40] Vaswani, A. et al. (2017). Attention is all you need. pages 5998–6008.

[41] Wang, Y., Zhang, L., and Chen, M. (2024). Generative language models with retrieval improve generation for automated short answer scoring.

[42] Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in test average model performance. *Clim. Res*, 30(1):79–82.

[43] Wolf, T. et al. (2020). Transformers: State-of-the-art natural language processing. In *in Proc. EMNLP: System Demonstrations*, pages 38–45.

[44] Zhang, Y., Wang, L., and Chen, M. (2024). Label-aware debiased causal reasoning for natural language reasoning. *AI Open*, 5:58–67.

[45] Zhu, Y., Chen, L., and Wang, X. (2025). Towards human-like grading: A unified llm-enhanced system for subjective question testing. *Front. Artif. Intell. Appl*, 392:1361–1368.