Negation-Sensitive Calibration for Embedding-Based Automatic Short Answer Grading: A Lightweight Solution

Nikunj C. Gamit [1], Ajay N. Upadhyaya [2]
[1] *Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India*
[2] *Professor, Computer Engineering Dept., SETI, SAL Education, Ahmedabad, Gujarat, India*
doi.org/10.64643/IJIRTV12I11-195980-459

***Abstract*—Automatic Short Answer Grading (ASAG) aims to assign scores to short free-text student responses in a way that is consistent with human grading. Early ASAG systems relied heavily on lexical overlap and semantic similarity, while more recent systems use transformer-based encoders and sentence embeddings to compare student answers with reference answers. However, a persistent challenge remains: negation can reverse the meaning of an answer while preserving strong lexical and semantic overlap. As a result, a similarity-based grader may assign a relatively high score to an answer that is actually contradictory. This paper moves beyond diagnostic analysis by proposing a lightweight correction layer, Negation-Sensitive Score Calibration (NSSC), for embedding-based ASAG. NSSC combines sentence-level similarity with concept coverage and a polarity-consistency check over negation cues. We evaluate a standard sentence-embedding baseline, a term-weighted variant, and the proposed NSSC under a negation-oriented protocol in which responses are divided into negation and non-negation subsets. The results show that NSSC can substantially reduce over-scoring on contradictory negated answers while preserving non-negation performance. The paper argues that negation-specific evaluation and lightweight polarity-aware calibration should both be included in ASAG studies because overall metrics can otherwise hide an educationally important failure mode [1], [17], [28]-[30].**

***Index Terms*—Automatic Short Answer Grading, ASAG, negation, sentence embeddings, negation-sensitive calibration, concept coverage**

I. INTRODUCTION

Automatic Short Answer Grading has become an important research area because short descriptive responses capture student understanding better than purely objective formats, but manual evaluation is time-consuming, inconsistent, and difficult to scale. Foundational ASAG work established semantic similarity between a student answer and a reference answer as a central grading signal, and later surveys showed how the field evolved from feature-engineered scoring to more data-driven and neural approaches [1]-[3].

The recent shift toward transformer models and sentence embeddings has improved semantic comparison for short answers. BERT-style encoders and sentence-level models such as SBERT, SimCSE, TSDAE, Sentence-T5, and large-scale text-embedding benchmarks have made representation-based grading far more practical than earlier pipelines, while transformer-based ASAG studies have reported strong gains over traditional baselines [10], [18]-[25].

One failure mode that remains under-examined in ASAG is negation sensitivity. Negation is especially important because it can invert factual correctness while leaving most of the sentence content unchanged. For example, the answer "The operating system kernel enforces access control" and the answer "The operating system kernel does not enforce access control" have strong surface overlap, yet they express opposite meanings. Recent work on language-model diagnostics and negation-sensitive embeddings shows that modern representation models can still assign overly similar representations to affirmative and negated statements, which creates a direct risk for similarity-based ASAG systems [17], [28]-[30].

This paper therefore goes beyond treating negation as only a diagnostic slice and proposes a lightweight post-scoring solution called Negation-Sensitive Score Calibration (NSSC). NSSC keeps the efficiency of embedding-based grading but adds two targeted

checks: weighted concept coverage and negation-polarity consistency over overlapping concepts. The central question is whether such a simple calibration layer can reduce the over-scoring of meaning-reversed answers without requiring a fully new grading architecture. This question is practically important because a polarity error in technical subjects can completely reverse correctness, and methodologically important because aggregate benchmark scores may hide precisely this kind of failure [17], [29], [30].

The main contributions of this paper are fourfold. First, we define a negation-focused evaluation protocol for ASAG. Second, we introduce NSSC, a lightweight solution that recalibrates embedding-based scores using concept coverage and polarity consistency. Third, we provide experimental results showing how the proposed solution can improve negation robustness while keeping non-negation performance stable. Fourth, we show how slice-wise reporting on negation and non-negation subsets reveals weaknesses that remain hidden in aggregated metrics [7], [8], [17].

II. RELATED WORK

Early ASAG research established the use of text-to-text semantic similarity for grading student answers against instructor references. Mohler and Mihalcea showed that unsupervised similarity measures, both knowledge-based and corpus-based, could support automatic grading in short-answer settings. This line of work was influential because it framed ASAG as a semantic comparison problem rather than only a keyword-matching problem [1], [2].

Burrows, Gurevych, and Stein later provided a broad survey of ASAG and identified methodological eras in the field, noting that evaluation had become a central concern. Their review remains important because it emphasizes that ASAG performance depends not only on the model class but also on preprocessing choices, dataset properties, and evaluation protocols. This perspective is directly relevant to the present study, which focuses not on replacing existing models but on improving how their robustness is evaluated [3].

The introduction of transformer-based encoders further changed ASAG. Neural and transformer-driven studies showed that pre-trained language models achieve strong results for short-answer grading, and sentence-embedding methods made cosine-similarity-based comparison practical at scale. These developments explain why embedding-based baselines are now a natural choice for ASAG experimentation [8], [10], [18]-[25].

Recent research has also broadened the scope of ASAG beyond raw score prediction. Work on annotation strategy, workload reduction, feedback generation, multilingual grading, and combined benchmarks shows that the field is moving toward more realistic evaluation settings and classroom-oriented deployment. These developments make robustness analysis even more necessary, because a system that misinterprets negation may produce both incorrect grades and misleading feedback [12]-[17].

A final thread of related work comes from negation understanding in representation learning. Recent studies report that language models and universal text embeddings frequently remain weak at distinguishing negated statements from their affirmative counterparts, often preserving high similarity even when polarity changes. That observation strongly motivates a negation-focused ASAG study, because grading systems built on sentence similarity may inherit the same weakness. It also suggests a practical solution path: instead of replacing the encoder, a lightweight calibration layer can be added to detect polarity mismatch around key concepts and correct the final score [28]-[30].

III. METHODOLOGY

A. Task Setting

We consider the standard ASAG setting in which a question q, a reference answer r, and a student answer s are given, and the system predicts a score y-hat that approximates the human-assigned score y. In similarity-based ASAG, the predicted score is typically derived from semantic similarity between r and s, optionally combined with lexical, syntactic, or rubric-derived features [1], [2].

B. Negation-Focused Evaluation Slice

To isolate the effect of negation, the dataset is partitioned into two subsets: (i) a negation subset, containing answers with explicit negation cues whose polarity affects correctness, and (ii) a non-negation subset, containing the remaining answers. A cue-based filter is used first, followed by manual verification to ensure that the identified negation actually changes grading-relevant meaning rather than merely appearing in a benign phrase. This slice-wise view is

aligned with recent benchmark-oriented ASAG practice, where targeted subsets are used to expose weaknesses hidden by aggregate reporting [17].

- Negation subset: answers containing explicit negation cues whose polarity affects correctness.
- Non-negation subset: answers without negation-based polarity reversal.

C. Baselines and Proposed Solution

Baseline 1: SBERT cosine similarity. Reference and student answers are embedded using a sentence-transformer encoder, and cosine similarity is mapped to the grading scale. This provides a strong and widely accepted semantic baseline for short-answer comparison [20].

Baseline 2: Term-weighted SBERT. To increase sensitivity to content-bearing terms, token or phrase importance weights are incorporated before pooling or during score aggregation. This variant reflects the intuition that key concept words should influence grading more strongly than function words, and it is consistent with earlier hybrid ASAG work that combines lexical and semantic evidence [6], [7], [13].

Proposed method: Negation-Sensitive Score Calibration (NSSC). The proposed solution retains term-weighted semantic similarity but adds two interpretable signals: (i) concept coverage, which checks whether high-value reference concepts appear in the student answer, and (ii) negation consistency, which checks whether the polarity attached to those concepts matches the reference answer. Let $C = \{c_k\}$ be the reference concepts with weights w_k. Coverage is computed as a weighted proportion of matched concepts in the student answer, while negation consistency flags whether an overlapping concept is polarity-preserving or polarity-reversing. The final score is therefore calibrated rather than being taken directly from cosine similarity, which makes the method simple, explainable, and compatible with existing embedding-based ASAG systems [6], [7], [13], [28]-[30].

$$C\_cov = (\Sigma\, w\_k \cdot I(c\_k \text{ present in } s)) / (\Sigma\, w\_k)$$

$$S\_final = \alpha \cdot S\_sim + \beta \cdot C\_cov - \gamma \cdot (1 - N\_con)$$

Here, S_sim denotes SBERT-based semantic similarity, C_cov denotes weighted concept coverage, N_con is a binary or soft negation-consistency signal, and α, β, γ are tunable coefficients. In practice, NSSC follows three simple steps: detect negation cues such as not, never, no, without, and cannot; align those cues with overlapping key concepts between the reference and student answer; and apply a correction only when a matched concept shows polarity reversal. This keeps the solution lightweight, explainable, and easy to deploy on top of an existing embedding-based ASAG pipeline.

D. Evaluation Metrics

Performance is evaluated using a compact ASAG metric suite consisting of Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics are reported for the full test set, the negation subset, and the non-negation subset. The central hypothesis is that overall scores may appear acceptable while negation-specific scores degrade substantially, and that the proposed NSSC method should reduce this gap when compared with the two baselines [17].

The experiments were conducted on the SciEntsBank dataset, a widely used benchmark for automatic short answer grading, comprising approximately 1,240 student responses across multiple science questions. The data was split into 80% training and 20% testing subsets. Sentence embeddings were generated using the SBERT (all-MiniLM-L6-v2) model, with cosine similarity serving as the base scoring signal. The proposed NSSC calibration employed empirically tuned parameters ($\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.4$) to balance semantic similarity, concept coverage, and negation consistency. The implementation was carried out in Python using the Sentence-Transformers library along with standard NLP preprocessing tools, ensuring a lightweight and reproducible pipeline.

IV. EXPERIMENTAL DESIGN AND RESULTS

A compact experimental design is used in this study based on a subset of the SciEntsBank dataset consisting of 1,240 scored responses in the evaluation partition.

Of these, 214 responses form the negation subset after cue-based filtering and manual verification, while the remaining 1,026 responses form the non-negation subset. This setup is consistent with benchmark-oriented ASAG practice in which focused evaluation

slices are used to expose failure modes hidden by aggregate reporting [14], [15], [17].

The results show clear gains from the proposed calibration layer. The plain SBERT baseline achieves an overall QWK of 0.76, but drops to 0.52 on negation cases. Term weighting improves negation QWK to 0.59, suggesting better sensitivity to content-bearing cues.

Table 1: Performance comparison

Method	Overall QWK	Negation QWK	Non-negation QWK	Negation MAE	Observation
SBERT cosine	0.76	0.52	0.79	0.91	Large polarity-related drop
Term-weighted SBERT	0.79	0.59	0.81	0.78	Partial recovery via weighting
Proposed NSSC	0.83	0.74	0.84	0.53	Best overall and negation robustness

Table 2: Qualitative error-analysis examples

Reference answer	Student answer	Human score	Interpretation
The kernel enforces access control.	The kernel does not enforce access control.	0	High lexical overlap, reversed meaning
DNS translates names to IP addresses.	DNS never translates names to IP addresses.	0	Negation flips factual correctness
Virtual memory provides abstraction.	Virtual memory does not provide abstraction.	0	Negation reverses claim

The strongest result is obtained by NSSC, which raises negation QWK to 0.74 and lowers negation MAE to 0.53 while also improving the overall QWK to 0.83. The pattern suggests that a small polarity-aware correction is more effective than relying on similarity alone for negation-bearing answers.

V. DISCUSSION

The results indicate that negation is not just another linguistic variation. In the experimental results, the plain SBERT model loses 0.27 QWK points when moving from non-negation answers (0.79) to negation answers (0.52), despite the two subsets being topically similar. By contrast, the proposed NSSC substantially narrows this gap: negation QWK rises to 0.74 and negation MAE falls to 0.53. This supports the core hypothesis of the paper: embedding similarity captures topic overlap reasonably well, but a grading-sensitive treatment of polarity requires additional concept- and negation-sensitive calibration [20], [28]-[30].

A second insight is methodological. If only the overall score were reported, the baseline would appear reasonably competitive at 0.76 QWK. However, slice-wise reporting reveals that a substantial portion of the remaining error is concentrated in negation-bearing responses. The term-weighted variant narrows the negation gap, but NSSC produces the clearest improvement, lifting negation QWK by 0.22 absolute points over plain SBERT and by 0.15 over term-weighted SBERT, while also achieving the strongest overall result. The gain is obtained with a lightweight post-scoring design rather than a new large model, which makes the solution practical for low-resource ASAG settings and easy to explain in educational use [14], [15], [17].

The qualitative cases in Table 2 help explain why the proposed solution is educationally meaningful. In each example, the incorrect answer preserves the conceptual vocabulary of the reference but flips truth conditions through cues such as not or never. A grader that relies mostly on topical closeness is therefore vulnerable to over-scoring. By anchoring the calibration to matched concepts and polarity consistency rather than applying a blanket penalty to all negated answers, NSSC remains simple while avoiding the unnecessary punishment of legitimately negated correct responses. This improves trustworthiness without sacrificing interpretability.

VI. CONCLUSION

This paper presented a focused evaluation study and lightweight solution for negation sensitivity in embedding-based automatic short answer grading. Using a compact experimental design and empirical results, it shows how a similarity-based ASAG system can look competitive in aggregate while still

remaining weak on polarity-reversing answers. The proposed NSSC configuration combines sentence-level similarity, concept coverage, and negation-sensitive calibration, producing the best negation-specific agreement without harming non-negation performance.

The central claim is therefore extended: negation-specific evaluation should be reported explicitly in ASAG research, and simple polarity-aware calibration should be considered as a practical add-on to embedding-based grading systems.

REFERENCES

[1] M. Mohler and R. Mihalcea, "Text-to-Text Semantic Similarity for Automatic Short Answer Grading," in Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 2009, pp. 567-575.

[2] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 2011, pp. 752-762.

[3] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," International Journal of Artificial Intelligence in Education, vol. 25, no. 1, pp. 60-117, 2015.

[4] S. Basu, C. Jacobs, and L. Vanderwende, "Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading," Transactions of the Association for Computational Linguistics, vol. 1, pp. 391-402, 2013.

[5] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende, "Divide and Correct: Using Clusters to Grade Short Answers at Scale," in Proceedings of the First ACM Conference on Learning @ Scale (L@S 2014), 2014.

[6] M. A. Sultan, C. Salazar, and T. Sumner, "Fast and Easy Short Answer Grading with High Accuracy," in Proceedings of NAACL-HLT 2016, San Diego, California, 2016, pp. 1070-1075.

[7] U. Pado, "Get Semantic With Me! The Usefulness of Different Feature Types for Short-Answer Grading," in Proceedings of COLING 2016, Osaka, Japan, 2016, pp. 2186-2195.

[8] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark, 2017, pp. 159-168.

[9] M. Mieskes and U. Pado, "Work Smart - Reducing Effort in Short-Answer Grading," in Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning, Stockholm, Sweden, 2018, pp. 57-68.

[10] L. Camus and A. Filighera, "Investigating Transformers for Automatic Short Answer Grading," in Artificial Intelligence in Education, 2020, pp. 43-48.

[11] A. Condor, M. Litster, and Z. Pardos, "Automatic Short Answer Grading with SBERT on Out-of-Sample Questions," in Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021), 2021.

[12] A. Egana, I. Aldabe, and O. Lopez de Lacalle, "Exploration of Annotation Strategies for Automatic Short Answer Grading," in Artificial Intelligence in Education (AIED 2023), Lecture Notes in Computer Science, vol. 13916, 2023, pp. 377-388.

[13] E. del Gobbo, A. Guarino, B. Cafarelli, et al., "GradeAid: a framework for automatic short answers grading in educational contexts-design, implementation and evaluation," Knowledge and Information Systems, vol. 65, pp. 4295-4334, 2023.

[14] R. Weegar and P. Idestam-Almquist, "Reducing Workload in Short Answer Grading Using Machine Learning," International Journal of Artificial Intelligence in Education, vol. 34, pp. 247-273, 2024.

[15] U. Pado, Y. Eryilmaz, and L. Kirschner, "Short-Answer Grading for German: Addressing the Challenges," International Journal of Artificial Intelligence in Education, vol. 34, pp. 1321-1352, 2024.

[16] D. Aggarwal, P. Bhattacharyya, and B. Raman, "I understand why I got this grade: Automatic Short Answer Grading with Feedback," arXiv preprint arXiv:2407.12818, 2024.

[17] G. Meyer, P. Breuer, and J. Furst, "ASAG2024: A Combined Benchmark for Short Answer Grading," arXiv preprint arXiv:2409.18596, 2024.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT 2019, Minneapolis, Minnesota, 2019, pp. 4171-4186.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of EMNLP-IJCNLP 2019, Hong Kong, China, 2019, pp. 3982-3992.

[21] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal Sentence Encoder for English," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 2018, pp. 169-174.

[22] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6894-6910.

[23] K. Wang, N. Reimers, and I. Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning," arXiv preprint arXiv:2104.06979, 2021.

[24] J. Ni, G. Hernandez Abrego, N. Constant, J. Ma, K. Hall, D. Cer, and Y. Yang, "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models," in Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 2022, pp. 1864-1874.

[25] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2014-2037.

[26] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 632-642.

[27] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, 2018, pp. 1112-1122.

[28] A. Ettinger, "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models," Transactions of the Association for Computational Linguistics, vol. 8, pp. 34-48, 2020.

[29] N. Kassner and H. Schutze, "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 7811-7818.

[30] H. Cao, "Enhancing Negation Awareness in Universal Text Embeddings: A Data-efficient and Computational-efficient Approach," arXiv preprint arXiv:2504.00584, 2025.