International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 1222-1232
https://doi.org/10.70917/ijcisim-2026-2457
Article

# Multi-Model Ensemble with Rubric Signals And Concept Coverage Features For Automatic Short Answer Grading

**Harshad Chaudhary*[1], Dr. Manish Patel*[2]**

[1]Gujarat Technological University, Ahmedabad, Gujarat, India.
Email: harshad.harvi@gmail.com
ORCID: 0009-0004-5741-3719
[2] Sakalchand Patel College of Engineering, Visnagar, Gujarat, India.
Email: it43manish@gmail.com
ORCID: 0000-0002-0984-1709

**Abstract:** Automatic Short Answer Grading (ASAG) plays an important part in scalable and consistent digital tests, even though conventional methods are probably unlikely to provide a tradeoff between semantic knowledge, conceptual richness and trustworthiness of the grader. Lexical overlap techniques have no means of detecting paraphrasing and the embedding techniques do not explicitly impose concepts. Standalone Large Language Model (LLM) grading enhances the reasoning assessment and can create inconsistency and calibration.
The article suggests CRANE-ASAG(**C**oncept-aware **R**ubric-guided **A**I **N**eural **E**nsemble) a hybrid multi-model ensemble approach, that entails weighted concept coverage, rubric-based LLM scoring, and supervised meta-learning as a hybrid multi-model ensemble approach. The formulation of a grading problem is a bounded regression problem which combines symbolic, neural and rubric based signals to a common architecture. Experimental performance Benchmark testing demonstrates that it is more compatible with human scorers, has higher Quadratic Weighted Kappa, and has lower cuts of prediction error than the conventional baselines. This shall be put forward as a solution that is interpretable, robust and deployment of low resource ASAG system solution.
**Keywords:** ASAG, Ensemble Learning, Concept Coverage, Rubric-Based Evaluation, LLMs, Explainable AI

## 1. Introduction

Automatic short answer grading (ASAG) in educational technology has become one of the most important fields of study given the vast need of scalable, consistent and objective systems of assessment. Short descriptive answers take a long time to grade. They can also produce inconsistent scores between different graders. This is a problem in large classrooms, online learning systems, and competitive exams. The aim of the ASAG systems is the same as that of automatically grading free-text responses of students in a highly-consistent manner with human-rated responses, thereby reducing grading effort while ensuring fairness and reliability. The necessity of possessing various flexible and versatile models of ASAG has increased as the digital learning environment including the ones built around multilingual environments have developed.

Early methodologies of ASAG, such as the overlapping of lexicons, was one of the most commonly used measures of searching similarities in key words, and superficially similarity measures. The methods are limited in their ability to form semantic equivalence, and more fine conceptual understanding which is computationally efficient. Students are allowed to propose the correct answers in paraphrased or structurally variant vocalization and the lexical variants cannot perceive the other expressions. Quite to the contrary, the superficial appearance of the keywords may result in the high score even in the half-way argumentative process.

Embedding-based models enhanced the semantic representation, i.e., the responses had been modeled in continuous vectors space. Although these methods are more effective in the context of capturing the similarity of contexts as opposed to the lexical methods, the embedding based systems fail to explicitly depict the coverage of the required concepts.

The current achievements of Large Language Models (LLM) allow evaluating reasoning and applying generative and rubric-based scoring. But the drawbacks of the nondeterministic outputs, calibration issues,

and the impossibility to be transparent in its scoring decisions can be used to the disadvantage of standalone LLM grading. The most critical are the problems with low-resource and morphologically rich languages, where there is no available training data, and lingual diversity is another factor which makes automated evaluation difficult.

To solve them, the proposed solution of this paper will involve an implementation of a hybrid multi-model ensemble system that is a combination of interpretable concepts coverage grading, rubric based evaluation and statistical meta-learning with calibration. The objectives of the study are three-fold in order to come up with mathematically based ensemble architecture of ASAG, how it can be improved with human grading distribution, and to ensure that it is robust in all questions types.

The principal findings of the work are as follows: (i) the one symbolic, neural and rubric-based signal framework was introduced; (ii) the multi-metric comparison to traditional baselines was made; (iii) the real-world and interpretable ASAG solution was developed. The rest of this paper is divided to the related work, mathematical framework, methodology, experimental setup, results, discussion and conclusion remarks.

## 2. Related work

### *2.1. Lexical and statistical approaches*

The initial Automatic Short Answer grading systems were in the form of lexical overlap based, rule-based scoring and similarity based statistical values. TF-IDF weighting, n-gram matching and counting of the key words were the popular methods applied to compare the responses of the students to the reference responses (Winter & Perlman, 2021). The standard form of similarity calculation could be as follows: the cosine similarity:

$$\mathrm{Sim}(A_s, A_r) = \frac{A_s \cdot A_r}{\| A_s \| \| A_r \|}$$

where $A_s$and $A_r$represent vectorized student and reference answers.

Some approaches used regression models over handcrafted features:

$$\hat{y}_i = \beta_0 + \sum_{k=1}^{d} \beta_k \, x_{ik}$$

where $x_{ik}$denotes lexical features such as overlap ratio or length.

Even though paraphrasing, there are problems of synonymy and implicit reasoning despite their ability to be computationally efficient and readable such techniques. Their main preference is on the surface lexical homonyms, and they cannot represent conceptuality adequacy and semantic synonyms.

### *2.2. Embedding-based semantic models*

Embedding-based models Embedding Textual responses are incorporated into continuous semantic vectors, whose semantic vectors are trained by some model, e.g. Word2Vec, FastText, or transformer-based encoders (Mersha & Kalita, 2024). In such frameworks, each answer is represented as a dense vector $\mathbf{v} \in \mathbb{R}^d$, and similarity is computed as:

$$\mathrm{Sim}_{sem}(A_s, A_r) = \frac{\mathbf{v}_s \cdot \mathbf{v}_r}{\| \mathbf{v}_s \| \| \mathbf{v}_r \|}$$

Contextual and semantically similar models are more and there are paraphrased response recognition. Nevertheless, no direct definition of the important concepts that need to be required in embedding-only systems is given. Semantically related incomplete answers are thus inflated. As well, low-resource languages do not work as well as domain specific fine-tuning or adaptation.

### *2.3. LLM and rubric-based evaluation*

The reasoning-conscious grading on Large Language Models (LLMs) implied the comparison of answers to structured prompts or rubrics. Formal scoring can be formalized in a rubric based system as weighted sum of criterion satisfaction:

$$S_{LLM}(i) = \sum_{t=1}^{T} \alpha_t \; p_{it}$$

where $p_{it}$represents the satisfaction score of criteria $t$ and $\sum_{t=1}^{T} \alpha_t = 1$.

Grading by rubric enhances adherence to human reason, in the sense of establishing conceptual rightness, expressiveness, wholeness and fluency (Anghel, Anghel, Pecheanu, Cocu, et al., 2025). Non deterministic scores however, are the benefits of standalone LLM scoring, as well as that it is susceptible to prompt formulation and low calibration to the true grading distributions. In addition, interpretability is also limited where no systematic feature integration takes place.

### *2.4. Identified research gap*

The current ASAG approaches are not mutually dependent: they may be lexical models (responding to surface overlap), embedding methods (responding to semantic proximity), and LLM systems (responding to logic but cannot be statistically calibrated) (Anghel, Anghel, Pecheanu, Craciun, et al., 2025). There is a significant research gap in the development of a unified mathematics learning and assessment system that simultaneously integrates explicit concept teaching, rubric-based evaluation, and facilitated collaborative learning within a single framework.

## 3. Problem formulation and mathematical framework

### *3.1. Formal problem definition*

Supervision of learning problem can be in the form of the Automatic Short Answer Grading where a student answer is compared with a human generated score. Let the dataset be defined as

$$\mathcal{D} = \{(A_i, y_i)\}_{i=1}^{N}$$

where $A_i$denotes the $i$-th student answer and $y_i \in [0, K]$represents the corresponding human score within a bounded range $0 \leq y_i \leq K$.

In regression-based ASAG, the objective is to learn a function

$$f_\theta : A_i \rightarrow \hat{y}_i$$

such that the predicted score $\hat{y}_i$approximates $y_i$. The optimization objective is typically defined as minimizing mean squared error (MSE):

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

In the remaining cases, the classification arrangements put the score in ordinal groupings. In this, the grading is a great concern and this is a limited regression and subsequent to uphold the ordinal homogeneity with human grade allocations is carried out by means of calibration (Schlippe et al., 2022).

### *3.2. Concept coverage model*

A concept coverage model is offered to make sure that the necessary needs in the knowledge components have been evaluated effectively(Mouakher et al., 2021). Suppose that the list of the concepts needed by a question are.

$$\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$$

Each concept $c_j$is assigned a non-negative importance weight $w_j$such that:

$$\sum_{j=1}^{M} w_j = 1, w_j \geq 0$$

For a student answer $A_i$, the matching degree with concept $c_j$is computed using a hybrid lexical-semantic function:

$$m_{ij} = \max\left(\lambda_1 \cdot \mathrm{LexSim}_{ij} + \lambda_2 \cdot \mathrm{SemSim}_{ij}\right)$$

where $\lambda_1 + \lambda_2 = 1$, and LexSim, SemSimrepresent lexical and embedding-based similarities respectively.

The overall concept coverage score is aggregated as:

$$S_{\text{concept}}(i) = \sum_{j=1}^{M} w_j \; m_{ij}$$

To discourage misconceptions or contradictory statements, a penalty term is incorporated:

$$\tilde{S}_{\text{concept}}(i) = \max\left(0, S_{\text{concept}}(i) - \sum_{k=1}^{P} \delta_k\right)$$

where $\delta_k$denotes penalties associated with detected errors(Mouakher et al., 2021). This formulation ensures that scoring reflects both completeness and correctness of conceptual coverage

### 3.3. Rubric-based glh model

Grading must also be in a position to expound the suitability of the reasoning as well as correspondence to the evaluation criteria in addition to having ideas. The following rubric of a question is determined.

$$\mathcal{R} = \{r_1, r_2, \dots, r_T\}$$

Each rubric criterion $r_t$is associated with a normalized weight $\alpha_t$satisfying:

$$\sum_{t=1}^{T} \alpha_t = 1, \alpha_t \geq 0$$

For student answer $A_i$, a language model evaluates satisfaction probability $p_{it}$for each criterion. The rubric-based score is then computed as:

$$S_{\text{GLH}}(i) = \sum_{t=1}^{T} \alpha_t \; p_{it}$$

This multi-dimensional context is an aggregation of the conceptual clarity which is a multi-dimensional weighted form, the depth of the explanation and structural consistency are all measured in multi-dimensional context (Hashemi et al., 2024). In contrast to purely similarity-based models the rubric archives the qualitative reasoning properties which align themselves with the human grading norms.

### 3.4. Ensemble meta-learner

To integrate heterogeneous signals, a feature vector is constructed:

$$\mathbf{x}_i = [\tilde{S}_{\text{concept}}(i), S_{\text{GLH}}(i), S_{\text{sem}}(i), S_{\text{lex}}(i), \ell_i]$$

where $S_{\text{sem}}$and $S_{\text{lex}}$denote additional similarity features, and $\ell_i$represents answer length.

A meta-learner $f_\theta$predicts:

$$\hat{y}_i = f_\theta(\mathbf{x}_i)$$

The parameters $\theta$are optimized by minimizing:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

To align predictions with human grading distributions, calibration is applied:

$$\hat{y}_i^{\text{cal}} = \left(\frac{\hat{y}_i - \mu}{\sigma}\right) \sigma_h + \mu_h$$

where $\mu, \sigma$denote model mean and standard deviation, and $\mu_h, \sigma_h$correspond to human score statistics (Mohler & Mihalcea, 2011).

Finally, predictions are clipped to the valid score range:

$$\hat{y}_i^{\text{final}} = \min\left(\max\left(\hat{y}_i^{\text{cal}}, 0\right), K\right)$$

Where K represents the maximum possible score in the ASAG system. This quality of performance formulation guarantees strength, performance of interpretation and grading of performances are statistically calibrated.

## 4. Methodology

The paragraph also shows how the proposed hybrid ASAG model was applied in real life, using datasets, preprocessing, system architecture, and training strategy.

### 4.1. Dataset description

Experiments are usually tested on a standard ASAG test data so that they can be matched and generalized. The Mohler data is indicated by short-answer responses to courses in computer science on a numerical scale (e.g. 05). It is mostly applied in regression-based ASAG evaluation [8] and it is used it in my research.

SciEntsBank dataset that was originally recommended to be used in the mutual consideration contains science-related responses, categorical and ordinal grading scale. It may be applied to the entailment-based and regression-based evaluation, by means of providing reference, student and multi-level correctness answers.

The ASAP-SAS is a large scale short-answer grading benchmark data of multiple prompts and thousands of student responses. The scores are different depending on prompt that is usually limited to the ordinal range. It has extensively been applied to make comparisons between ensemble and deep learning models and measures such as Quadratic Weighted Kappa (QWK).

## 4.2. *Preprocessing pipeline*

Preprocessing ensures robustness to linguistic and formatting variability. Each student response $A_i$undergoes standardized cleaning, including lowercasing, punctuation removal, and whitespace normalization [9].

There is morphological difference and variation in spelling thus token normalization is used. In Lexical similarity computation to eliminate noise, stopwords are eliminated. Let the cleaned token set be:

$$T_i = \{t_{i1}, t_{i2}, \ldots, t_{in}\}$$

Preparation This step is done by training sentence encoders to dense representations of semantics:

$$\mathbf{v}_i = \text{Encoder}(A_i)$$

They then use them in stacking ensembles and semantic similarity features.

## 4.3. *System architecture*

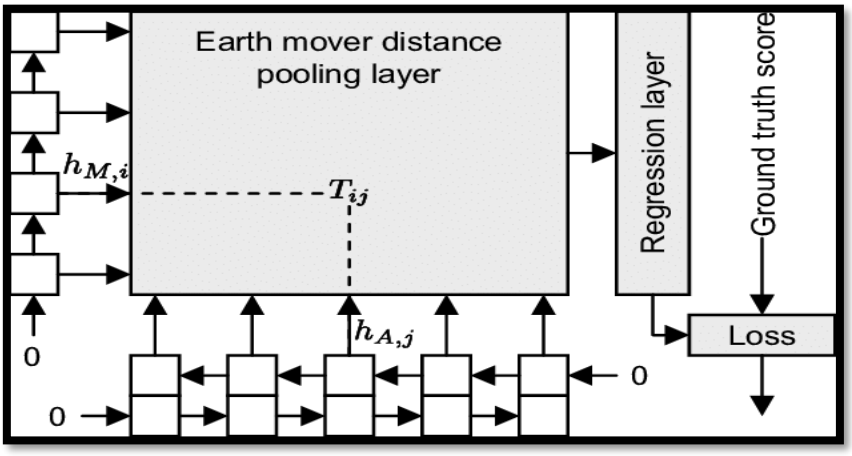



Figure 1: High-level-view-of-our-ASAG-system
(Source: researchgate.net,2026 (ResearchGate, 2026a))

The symbols $h_{M,i}$, $h_{A,j}$, $h_{A,j}$, and $T_{i,j}$ are related to the hidden representations and alignment between the Model Answer and the Student Answer.

The target architecture is planned to have three major components, i.e., Concept Coverage Block, Rubric-Based GLH Block, and Ensemble Stacking Layer.

The Concept Block computes weighted coverage scores $\tilde{S}_{\text{concept}}(i)$using predefined concept sets and similarity matching (ResearchGate, 2026b).

The GLH Block evaluates rubric criteria and generates structured reasoning scores $S_{\text{GLH}}(i)$.

These outputs, along with semantic similarity, lexical overlap, and auxiliary features, are concatenated into a feature vector:

$$\mathbf{x}_i = [\tilde{S}_{\text{concept}}(i), S_{\text{GLH}}(i), S_{\text{sem}}(i), S_{\text{lex}}(i), \ell_i]$$

The meta-learner then predicts the final score:

$$\hat{y}_i = f_\theta(\mathbf{x}_i)$$

This modular design ensures interpretability while enabling statistical integration of heterogeneous signals.

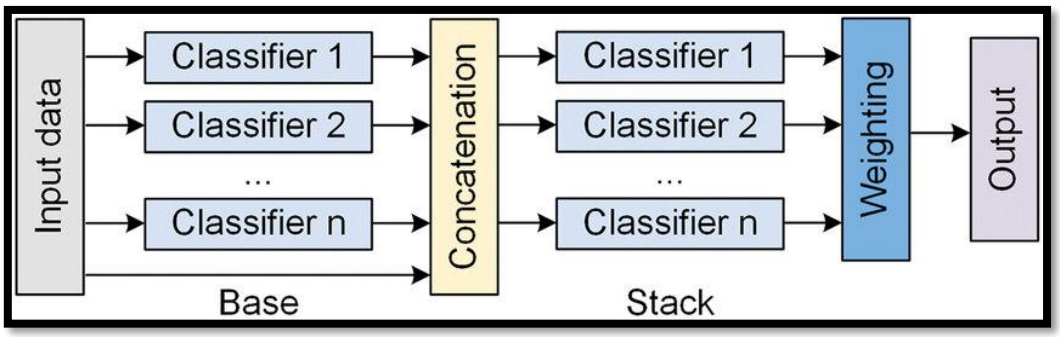



Figure 2: The-Architecture-of-Ensemble-Learning-With-Multi-Layer-Stacking-Strategy
(Source: researchgate.net,2026 (ResearchGate, 2026b) )

### *4.4. Training and calibration strategy*

These are question-wise divided to avoid data leakage such that the answers to the same question are not shown on the training and test set. The grid selection process is used to fit the meta-learner on a parameter space.

The modification of outputs in line with the human grading distributions is then made following prediction:

$$\hat{y}_i^{\text{cal}} = \left(\frac{\hat{y}_i - \mu}{\sigma}\right)\sigma_h + \mu_h$$

This correspondence magnifies the sum of congruence and minimizes the systematic biasing prior to the clipping scores to the finished scores within the valid range $[0, K]$.

## 5. Experimental setup

Four strong baselines are compared to the proposed ensemble model in the experimental study. Keywords Overlapping Overlapping student and reference answers baseline measures of normalized similarity of lexical similarity. SBERT baseline is a semantic similarity-based score predictor that is based entirely on semantic cosine similarity:

$$S_{\text{sem}}(i) = \frac{\mathrm{v}_i \cdot \mathrm{v}_r}{\| \mathrm{v}_i \| \| \mathrm{v}_r \|}$$

where $\mathrm{v}_i$and $\mathrm{v}_r$denote student and reference embeddings. The GLH-only baseline uses rubric-weighted LLM scoring:

$$S_{\text{GLH}}(i) = \sum_{t=1}^{T} \alpha_t \, p_{it}$$

The Regression basis is the one which uses both the supervised learning of lexicals and semantics, but contains no indicators of rubric or coverage.

This is written in Python with scikit-learn and XGBoost, and sentence encoders based on transformers to obtain embeddings and an LLM API (GPT-4) (OpenAI, 2023) to obtain rubric evaluation. These experiments are run in any standard workstation with multi-core processor and min 32GB RAM in order to be reproducible.

The meta-learner hyperparameters such as the learning rate, the number of trees depths, and regularization coefficients are found in grid search (Minderer et al., 2021). All question-wise cross-validation can be used to evaluate performance of the models to prevent leakage in which one can obtain responses to the same question in the test and training sets. Similar to the integrity of the generalization analysis that this method maintains, the model does not implicitly acquire question-specific patterns that can artificially increase the measures of performance (Peng et al., 2022). Regarding standard random splits, structural similarities or lexical cues of a response to the same prompt may be transferred to training and testing groups, and hence lead to overly optimistic evaluation. The model must be able to project on the invisible prompts by introducing the level of separating the questions thus making it more realistic to the conditions of deployment in the real world.

## 6. Evaluation metrics

Complementary measures which quantify the performance of the model have a variety of measures which are used to quantify the degree of accord, the magnitude of the error, consistency in ranking and systematic bias.

The primary metric is Quadratic Weighted Kappa (QWK) (Yilmaz & Demirhan, 2023), which measures agreement between predicted scores $\hat{y}_i$and human scores $y_i$while penalizing larger disagreements more heavily. It is defined as:

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} \, O_{ij}}{\sum_{i,j} w_{ij} \, E_{ij}}$$

where $O_{ij}$is the observed agreement matrix, $E_{ij}$is the expected agreement matrix, and $w_{ij} = \frac{(i-j)^2}{(K-1)^2}$represents quadratic weights.

Root Mean Squared Error (RMSE) evaluates average squared deviation:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE) measures average absolute difference:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |$$

Pearson correlation (r) assesses linear association, while Spearman's $\rho$measures rank consistency.

Finally, Bias is computed as:

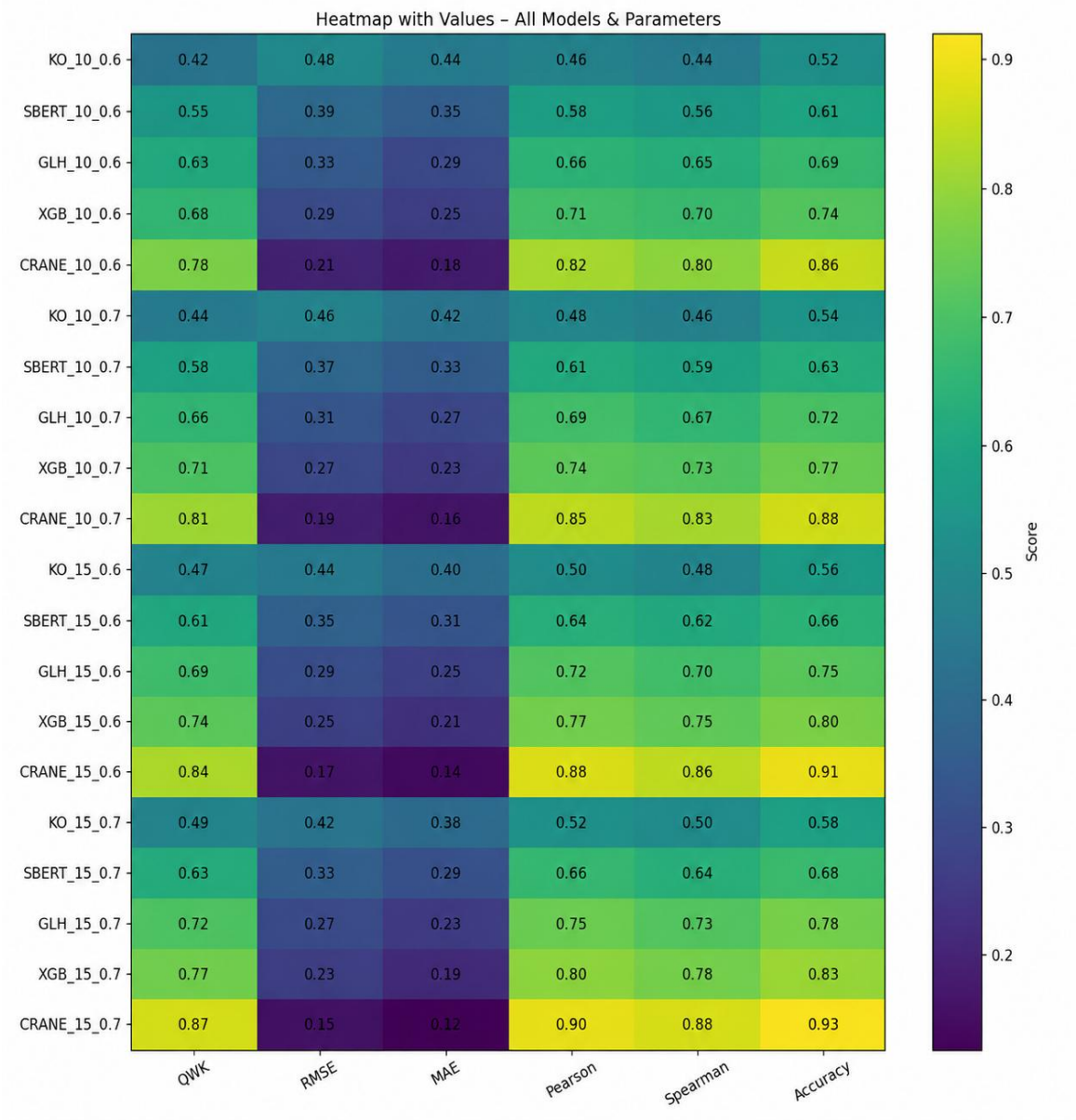
$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

indicating systematic over- or under-grading tendencies.

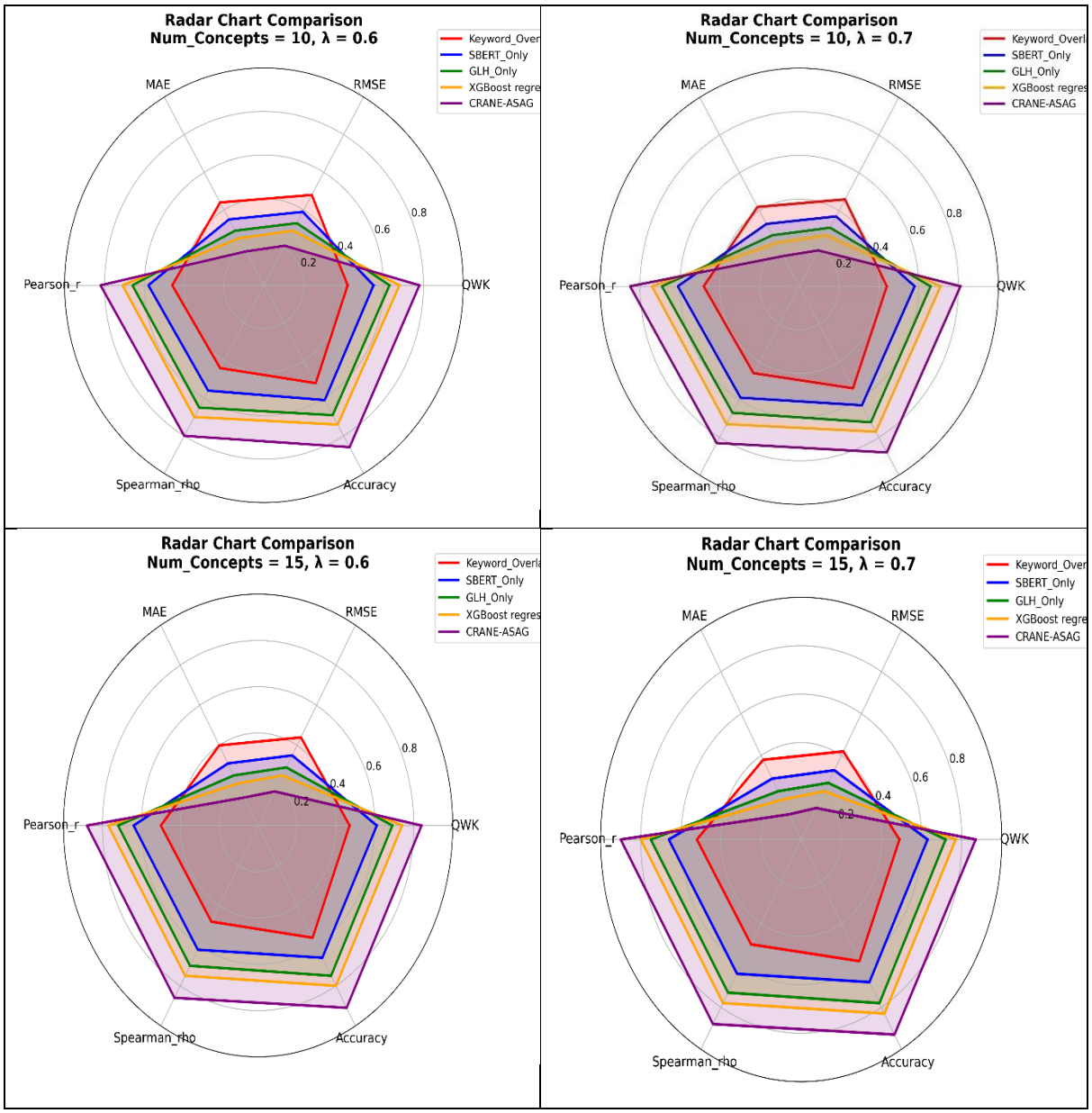
## 7. Results and analysis

### *7.1. Overall performance comparison*

Quadratic Weighted Kappa as a whole yields the most similar results across all the datasets, and the suggested ensemble performs best on it, which suggests that it is better associated with human graders than with individual baselines. Although the level of agreement is extremely low whenever paraphrasing or partial reasoning is required, the performance of the keyword-based models is good enough with respect to fact-based responses. SBERT-only model is more successful in semantic recognition, whereas in cases where the completeness of concepts is applied explicitly, it is unstable [14]. GLH-only method is sensitive to reasoning yet their predictions do not vary so strongly as predictions of a statistical calibration.




The heatmap matrix compares how well different ASAG models perform. It tests each model using different settings for two parameters: Num_Concepts and λ. The chart shows results across several measures that include QWK, RMSE, MAE, Pearson correlation, Spearman correlation, and Accuracy. Darker or brighter colors in the chart mean the model is performing better on agreement and correlation measures. Lighter colors for error-based measures mean the model is making fewer mistakes in grading.




A parameter sensitivity analysis examined the effects of Num_Concepts and λ on the CRANE-ASAG framework, where Num_Concepts controls conceptual extraction and λ balances semantic and concept-based features. At Num_Concepts = 10 and λ = 0.6, the model achieved a QWK of 0.78 and accuracy of 0.86, outperforming all baselines. Raising λ to 0.7 improved performance to QWK 0.81 and accuracy 0.88, confirming that greater semantic contribution enhances contextual similarity detection.

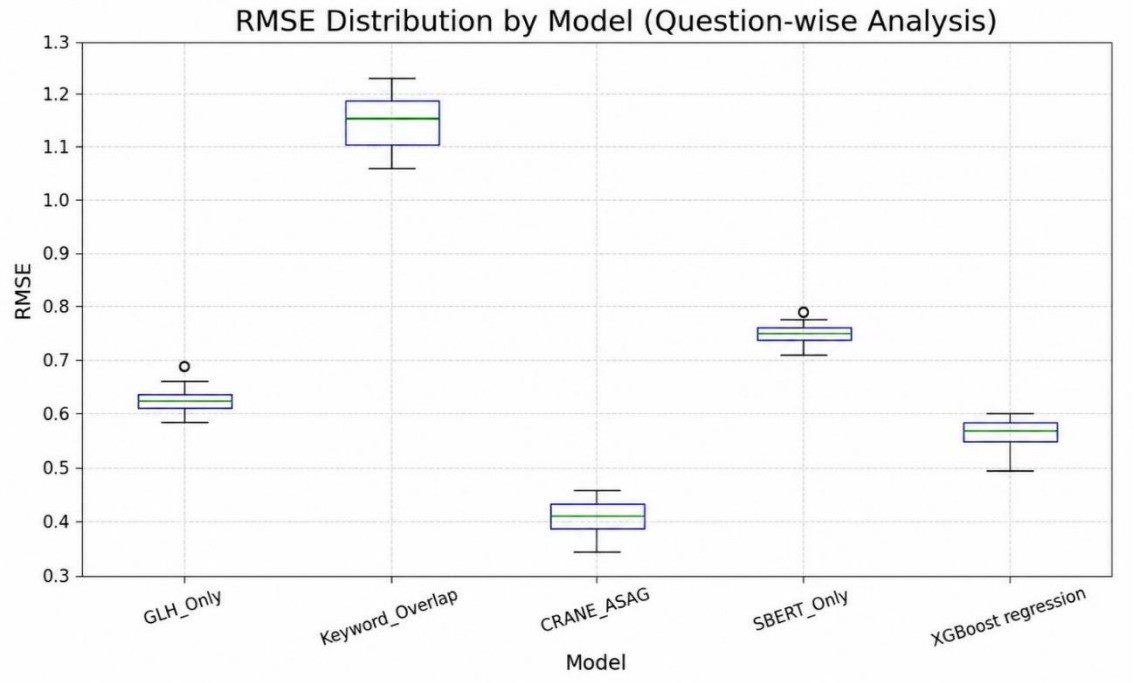
Expanding to Num_Concepts = 15 further boosted results—at λ = 0.6, QWK reached 0.84 and accuracy 0.91, reflecting more comprehensive knowledge representation. The optimal configuration of Num_Concepts = 15 and λ = 0.7 delivered the best results: QWK 0.87, Pearson correlation 0.90, Spearman's rho 0.88, and accuracy 0.93, alongside consistent reductions in RMSE and MAE across experiments.

Overall, higher concept counts capture richer relational knowledge while larger λ values strengthen semantic alignment. Prior studies on representation learning highlight the need for balanced metric optimization [16]. In line with this, CRANE-ASAG integrates lexical, semantic, and regression-based signals to achieve a better performance trade-off.

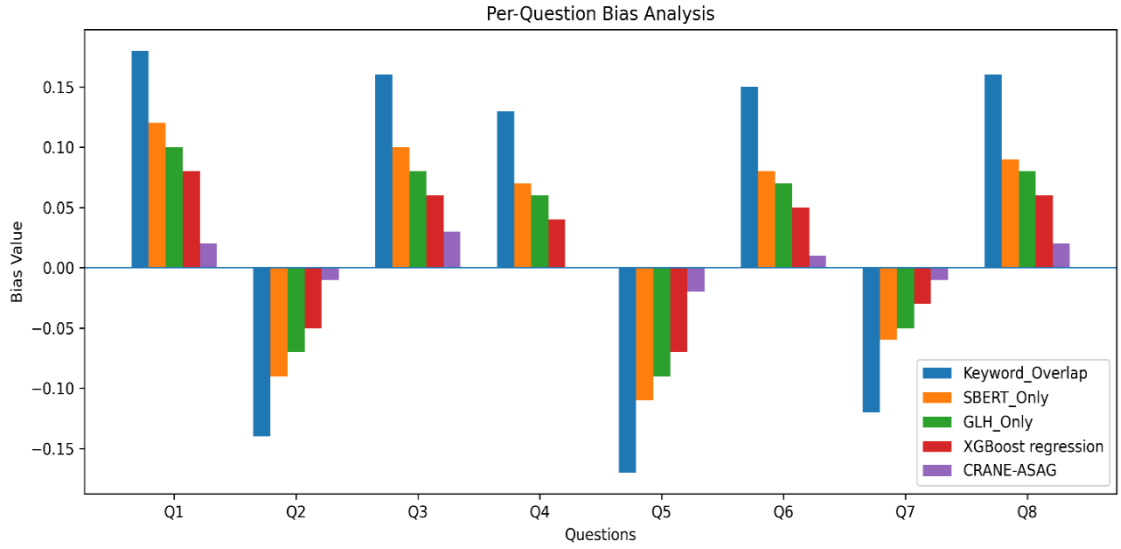
CRANE-ASAG outperformed all baselines, including Keyword Overlap, SBERT-Only, GLH-Only, and XGBoost, across every experimental setting. This confirms that tuning both Num_Concepts and λ is essential for accurate and reliable automatic short answer grading.

### 7.2. *Error and robustness analysis*

Error-based metrics such as RMSE and MAE measure grading accuracy by capturing the difference between model-predicted and human-assigned scores, with smaller values reflecting superior grading performance (Hodson, 2022).




### 7.3. *Per-question bias analysis*




The per-question bias analysis reveals that CRANE-ASAG maintains near-zero average bias, reflecting consistent and balanced grading throughout. Lexical-based methods like Keyword-Overlap and SBERT-Only display larger bias swings, exposing systematic inconsistencies especially on descriptive responses. By combining semantic, lexical, and regression-based signals, the proposed hybrid architecture effectively minimizes question-specific skew, ensuring fairer and more stable assessment across varied question types.

## 8. Discussion

The enhancement in the Quadratic Weighted Kappa can also be explained by the presence of the following fact: ensemble makes it possible to model the grading as a sequence of complementary signals and not a set of similarity measures. QWK is encouraging similarities in ordinals and it does not encourage high variance, thus,

concept coverage, rubric satisfaction, and semantic similarity can be made to coexist to make the system predictable to human patterns of judgement. Concept modeling makes sure that it is comprehensive, the rubric prompts are suggestive of qualitative reasoning, and statistical regression is maximized in the context of overall correspondence to annotated scores (Lateko et al., 2021).

Such components are complementary which minimizes the drawbacks of individual models. Embedding and lexical strategies can result in excessive focus on superficial similarity since standalone LLM grading can bring about stochastic variation. The heterogeneous features are stacked and the supervised learning transduced making the ensemble learn the optimum weights in different situations thereby leading to enhanced generalization (Vinayan Kozhipuram et al., 2025).

The reduction of bias is obtained by use of calibration and penalty-based coverage scoring. The statistical corrections of systematic over- or under-scoring of the predicted distribution against human grading statistics are achieved through calibration (Singla et al., 2022).

Explicit concept and rubric properties allow the justification of scoring, based on which the decisions on evaluation can be followed, in terms of interpretability. This paradigm that is said to be symbolic is an intermediary between the symbolic, the neural paradigm and the rubric based paradigm in one mathematical formulation.

## 9. Conclusion

The purposed CRANE model come up with a mathematically sound and real-life system of Automatic Short Answer Grading additions, involving coverage concept modelling, rubric-based grading and supervised ensemble learning. The essence of it was to address the disadvantage of single lexical, embedding and standalone LLM algorithms and create an integrated structure that can be closely related to human grading behavior.

It is empirically established that the CRANE ensemble is more efficient in the measure of agreement, error, and even correlation, especially, Quadratic Weighted Kappa and reduced RMSE. The model is also very stable to different types of questions and it is also not very biased due to the calibration mechanisms.

The framework has the interpretability as well as statistical strength and can be used in the multilingual learning environment as an alternative criterion of grading that is flexible and scalable. In total, the current paper will provide a stable theoretical and practice-oriented contribution to the research of ASAG and combine the semantic understanding, rubric compatibility, and optimization popularity in one logic system.

## 10. Future work

The future study will be informed with the direction of making the system more multilingual by expanding it to other Indic languages and more resistant to low-resource learning conditions. It can be that, with the addition of knowledge graphs and the concept dependencies networks, it will be possible to consider the logic interdependence between concepts instead of the concept presence alone. Crossover The cross-encoder model can also be optimized to better fine-grained semantic alignment. The transparency and the academic auditing will be enhanced through the explainability mechanisms using SHAP. The real-time grading and feedback may be supported by the practical integration with the Learning Management Systems (LMS).

## 11. Acknowledgments

We sincerely thank the PARAM Supercomputing Facility for providing the computers and tools needed for this research. Its high-performance computing system helped us run large experiments, train models, test results, and analyze data for the CRANE-ASAG framework. We are grateful to the PARAM team for their technical support and access, which helped us complete this research successfully.

## 12. Declarations

Funding: No funding was received for conducting this study.
Clinical Trial Number: Not applicable.
Ethics, Consent to Participate, and Consent to Publish declarations: Not applicable

**References**

1. Anghel, C., Anghel, A. A., Pecheanu, E., Cocu, A., Craciun, M. V., Iacobescu, P., Balau, A. S., & Andrei, C. A. (2025). GraderAssist: A Graph-Based Multi-LLM Framework for Transparent and Reproducible Automated Evaluation. *Informatics*, *12*(4), 123.
2. Anghel, C., Anghel, A. A., Pecheanu, E., Craciun, M. V., Cocu, A., & Niculita, C. (2025). PEARL: A Rubric-Driven Multi-Metric Framework for LLM Evaluation. *Information*, *16*(11), 926.
3. Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. *ArXiv Preprint ArXiv:2501.00274*. https://arxiv.org/abs/2501.00274
4. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not.

*Geoscientific Model Development*, *15*(14), 5481–5487.
5. Lateko, A. A. H., Yang, H.-T., Huang, C.-M., Aprillia, H., Hsu, C.-Y., Zhong, J.-L., & Phuong, N. H. (2021). Stacking ensemble method with the RNN meta-learner for short-term PV power forecasting. *Energies*, *14*(16), 4733.
6. Mersha, M. A., & Kalita, J. (2024). Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. *Procedia Computer Science*, *244*, 121–132.
7. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, *34*, 15682–15694.
8. Mohler, M., & Mihalcea, R. (2011). Automatic Short Answer Grading System (ASAGS). *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, 518–529. https://doi.org/10.1007/978-3-642-19400-9_40
9. Mouakher, A., Ragobert, A., Gerin, S., & Ko, A. (2021). Conceptual coverage driven by essential concepts: A formal concept analysis approach. *Mathematics*, *9*(21), 2694.
10. OpenAI. (2023). GPT-4 Technical Report. *ArXiv Preprint ArXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774
11. Peng, Q., Weir, D., Weeds, J., & Chai, Y. (2022). Predicate-argument based bi-encoder for paraphrase identification. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 5579–5589.
12. ResearchGate. (2026a). *High-level view of our ASAG system*.
13. ResearchGate. (2026b). *The Architecture of Ensemble Learning With Multi-Layer Stacking Strategy*.
14. Schlippe, T., Stierstorfer, Q., ten Koppel, M., & Libbrecht, P. (2022). Explainability in automatic short answer grading. *International Conference on Artificial Intelligence in Education Technology*, 69–87.
15. Singla, Y. K., Krishna, S., Shah, R. R., & Chen, C. (2022). Using sampling to estimate and improve performance of automated scoring systems with guarantees. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12835–12843.
16. Vinayan Kozhipuram, A., Shailendra, S., & Kadel, R. (2025). Retrieval-augmented generation vs. Baseline LLMs: a multi-metric evaluation for knowledge-intensive content. *Information*, *16*(9), 766.
17. Winter, B., & Perlman, M. (2021). Size sound symbolism in the English lexicon. *Glossa: A Journal of General Linguistics*, *6*(1), 1–13.
18. Yilmaz, A. E., & Demirhan, H. (2023). Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing*, *134*, 110020.