ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Generative AI-Enabled Automated Scoring Algorithms for College English Teaching

Xiaoyan Huang[1,*]

[1] School of General Education, Hunan University of Information Technology, Changsha, 410000, Hunan, China

SUMMARY: *Automated Scoring of college-English courses has gradually shifted from single-item essay evaluation to an overall evaluation system for Writing, Speaking, Translation, Reading Response, etc. Propose an automatic grading system based on generative artificial intelligence for multiple task-based college English courses. Combining task-conditioned language representation, rubric-aware generative reasoning, trait-level score fusion, ordinal score constraints, calibration correction, and uncertainty-based routing in the algorithm for human review. Organized a collection of 18,420 college English Responses from the following four instruction task types: 6,240 writing Samples, 4,180 Speaking Samples, 3,620 Translation Samples, and 4,380 Reading-Response Samples. For each of them, two trained raters scored according to the analytic rubric; their adjudicated labels were used for training and testing the scoring model. In the experiment, compared GAEAS with SVR, BiLSTM, BERT, T5, and GPT-4scoring in the same data split, scale of scores, and test criterion. In terms of all tasks, GAEAS achieved an approximation of the median QUCKW1=0. 903 and averaged a Pearson's Correlation Coefficient close to 0. 923; For the average value, it is roughly within the range of-47dBto~5dBandtheRMSEisappoximately±_QEUZS=-5%. GAEAS outperformed GPT-4o in terms of QWK score, achieved a reduction in RMSE and improved generalization performance (ECE) compared to the baseline; Inference time was reduced to 96ms per example. According to the score-band heatmap, most of these task-profanity combinations had a diagonal agreement rate in the range of 72%-84%. Through ablation experiments, it was confirmed that when using the combinations of ordinal loss, trait fusion, rubric verification and uncertainty routing simultaneously would yield optimal results. Further revealing through a three-dimensional hyperparameter response Surface shows that the prior-rubric weighing and calibration weights need to be changed together; The weakest validation Q-κ values are around 0.64 and 0.36, respectively. According to the results, it was found that with proper management of both Generative AI Scoring stability in College English Teaching Rubric Structure, Score Ordering Calibration and Human Review can all work together. The research also presents various results, including score-band alignment, component-level ablations, error-source distribution, and deployment trade-off situations of the algorithm to provide teaching tools instead of isolated models. According to the outcomes, automatic grading in college English needs to consider three dimensions: agreement, error magnitude, calibration costs, and post-evaluation expenses comprehensively. On the basis of this, a particular way for quick feedback along with teacher oversight remains as an option for some courses. Reproducibility of the protocol using fixed data splits and figure-level metric tables. This experiment is presented as a classroom-based empirical study rather than a review assertion; it uses the same held-out test set, fixes baselines, performs an explicit score-band analysis, and simulates a review threshold.*

*huangxiaoyanhuit@163.com
https://doi.org/10.65102/is2026776
Published: 30 April, 2026

KEYWORDS: *Generative artificial intelligence; automated scoring; college English teaching; rubric-aware assessment; calibration; uncertainty routing*

1 Introduction

Collegiate English teaching produces an abundant supply of measurable language artefacts. A regular teaching week may include short argumentative essays, oral presentations, translation exercises, passage-based responses, vocabulary-focused revisions, and formative reflections. Large-classroom teacher assessment of student performance; Explain reasons if applicable and make improvements in time by pointing out grammatical errors, providing feedback or advice on how to improve. The burden of course pressure is greater when there are combined courses of general English proficiency and disciplinary communications; any response may have issues in terms of content relevance, grammatical structure, coherence, lexical choice, pronunciations, etc. Human evaluation is still required to make decisions and hold responsible; However, there isn't much Time left for diagnosis or consultations during this period. This kind of design scheme provides a specific basis for evaluating the scores of college English tests according to existing score criteria.

Generative artificial intelligence has changed the technical conditions in which this problem is studied. Current work in language education now covers aspects such as text creation, feedback correction, pronunciation training exercises, reading prompts suggestions, and automatic grading of homework. Law's scoping review reported that GenAI applications in language teaching cover multiple instructional activities and raise new requirements for pedagogy-sensitive design [1]. A subsequent systematic review of empirical gen-AI studies found rapid accumulations in the research literature on learner interaction, feedback quality and classroom application [2]. The aforementioned results are all within the scope of evaluation for automated scoring and have direct applications in improving college English writing instruction. Therefore, it is necessary that a feasible algorithm should be able to accommodate the instruction cycle of teacher assignment, student revision and evaluation result feedback within an appropriate time frame.

Automated writing assessment systems are the most typical examples of this research. Relevant reviews of Grammarly, Pigai, Criterion, etc., show that automatic assessment can be used to monitor the accuracy of superficial expressions, revision behaviour, and teaching resources [3, 4]. In the classroom synthesis work, it was found that students' responses to automatic feedback vary according to different tasks, levels of proficiency and teachers' management. Clarify that a scoring function is not to be validated simply through a global agreement coefficient; In an English course, the same score may guide placement, formative feedback, peer review, or end-of-unit grading. Stable score distribution; Clear teacher evaluation criteria for the same aspect based on the rubric criteria; Sensitivity to changes reflected through particular item scoring results.

Research on automated essay-scoring algorithms has an older history. Early Systems included manualised feature, Lexical Index, Syntactic Indicators, Discourse Measure and Regression Models. The following methods are: Recurrent Neural Networks, Convolutional Networks, and Pre-trained Language Models. According to surveys by Ke and Ng, Ramesh and Sanampudi, among others, there is a shift in the direction of research towards representations learning; However, it still faces several problems including content-relevance, interpretability, prompt-transfer, etc. [5, 6] As these problems become more prominent during college English instruction due to diverse assessments. A student might generate a relatively fluid but unfirmly supported reading-response; an accurate but ideationally faulty translation; a well-structured oral report with poor intonation. A single-essay-centred mode fails to cover these situations

comprehensively.

Large-scale language models provide another possibility: by reading the criterion for evaluation, comparing it with a specific response, generating trait-based explanation materials, and replicating an expert's evaluation behaviour, etc. Using ChatGPT or similar models to automatically grade essays has shown relatively high correlation with human assessments but is inconsistent among different prompt situations, scoring sessions, etc. [7, 8]. According to the validity and reliability research results of English learner's writing, LLM score can be aligned with human judgment under certain circumstances but still highly sensitive to prompt Design and Score instruction. In direct classroom comparisons, ChatGPT scoring has also shown unstable alignment with experienced human raters under repeated scoring conditions [9]. The above results support using LLMs for scoring and also caution that treating a general generative model as an entire evaluation tool is inappropriate.

Currently, there are deficiencies in both the connection of generating meaning with evaluation criteria during Generative AI score setting; The model may produce plausible comments but have poor calibration of the scores. It can follow the criteria of good compositions, but has lower score expectations for average works. Can reach a high overall coherence degree while making significant errors on specific items related to detailed evidence gathering, language authenticity verification accuracy of discourse smoothness. Fairness work in automated essay scoring has further shown that high mean accuracy can coexist with subgroup-level bias, especially when demographic or proficiency variables are associated with writing competence [10, 11]. Therefore, a college English scoring algorithm needs to solve the problems of agreement, scores ordered, calibration, error management and manual verification simultaneously.

Recent hybrid methods tend to use this Design. Integrate an LLm-Scorinon Function with a separate Discriminative Predictor, Rubric-Characteristic/Calibration components to enhance Consistency and reduce Variance accordingly. Atkinson and Palma presented a combined use of LLMs to enhance automatic grading that showed some benefits when integrating geno-predictive signal with score prediction model. Systematic evidence at a broader scale shows that languages, in terms of tasks and answers generated by LLMs, has been widely applied; Reliability, Validity, Fairness, as well as Deployment Costs, are still core problems. [12] These researches offer some methodological references but fail to solve completely that problems existing at present: writing, speaking, translation and reading-response need to be scored on the same platform without separate task-rubric linkage.

This paper attempts to address this problem by creating a generative-ai-powered automatic-grading System for college English courses. The algorithm name is gaEAs. Each learner's response is treated as a task object with encoding of the language, rubric-based generation reasoning for generative scores, trait-level fusion, ordinal-score restriction, calibration adjustment, and uncertainty routing. A Design of four typical tasks found commonly in the College English Course: Writing Task Speaking Translation Reading-response. Measure the extent to which a generative-scoring algorithm maintains high agreement with human evaluators, produces stable score bands, has interpretability of trait evidence, and supports actual application deployment.

Ordinary university course's evaluation problem also has an impact of time. The students receive a grade for the next assignment in advance and cannot link their results to what they will learn later; Because teachers have focused on evaluating the final scores of students' explanation work. Students then see grades without enough evidence about which trait caused the score. An automated scoring algorithm that calibrates trait evidence to eliminate delays; However, it should not replace teachers' professional judgement with a sterile figure. To provide sufficient scoring evidence so that teachers can make a judgment of acceptance, modification

and rejection based on it.

The college English scenario is also task-variation-based and different from the high-stakes essay grading situation. Writing tasks generally include scores for reasons, organisational skills, usage of language, degree of vocabulary; Tasks for speaking include fluency and pronunciation cues that are not entirely obvious from the transcript. Translation tasks need to maintain the same meaning as before, but reading-responding needs specific materials in the text. The tasks have connections through the course objectives, but different standards of evaluation are applied in practice. If the scoring method cannot reflect this comprehensiveness at all, there would still likely lack much particularity to a teachers' present evaluation work;

Middle band response as a practical problem. Many college English students' answers fall into this category of being neither strong nor weak. They present a mix of outcomes; they are well-structured and grammatically correct, while the content is coherent in their answers orally; or translations with adequate grammar but unclear meanings. Generic scoring feedback generally cannot handle such ambiguous situations, as it converts complicated Rubric evidence into a one-size-fits-all requirement judgment. The algorithm that is more potent must be able to analyse and present data well when outputting results.

Calibrate it in time. A score of 7.0 with high certainty represents a different classroom connotation compared to the same score combined with high uncertainty. Teachers may utilise the first in routine feedback while storing the latter for further inspection. Without calibration, if the automatic scoring system uses reliable results to suppress possible dangers. The teaching context is not tolerable; if such serious violations occur, it may affect students' engagement in learning and perceptions by their peers to some extent. Therefore, confidence should be related to empirical agreement rather than an internal likelihood of the model.

Generative models also change the form of explanation. They can produce fluent comments that seem teacher-like, but fluency is not evidence of valid scoring. A comment may seem reasonable, but its corresponding number on the assessment criteria is mismatched. Therefore, the current study divides evidence reading from score determination separately. Generative functions identify relevant indicators for rubric evaluation; the output of the supervised scoring part constitutes an assessment result. The Design maintains an interpretive function for Generative AI and requires passing a learned constraint that has been trained using human ratings.

This research falls between the classroom demands and algorithms' controls in this boundary. It does not treat college English scoring as a general text classification problem. Scoring is treated as an organised task of evaluation; Response mode, Task instructions, Analytic rubric, Score bands, Human review thresholds all need to be satisfied for it to become valid. Also, this positioning can explain the figures in the paper. Method figures show the score setting, and result figures assess conformity, stability, calibration errors, deviation from expectations due to environment, etc., in conjunction with deployment costs.

Therefore, the specific research questions are as follows: Can a generative-aided scoring system enhance multi-task college English examination under the constraint of generative reasoning within the framework of rubric structures, ordinal scores' training, bias correction, and partial human supervision? A multiple-accuracy table cannot present it all at once. Require figures showing how the scores vary among Tasks, Score Bands, Modules, Confidence Levels, Classroom Review Constraints, etc. Below are the development of algorithms and evaluations based on this broad evidence threshold.

Another reason for focusing on college English is that the target of assessment is educationally mixed. Academic Writing, Oral Communication, Reading Comprehension, Translation Strategy training in the same course. The learning objectives are interrelated but cannot be simplified as an individual latent language capability. A good model in essay writing

may not perform well when evaluating whether a reading response is correct or if a translation has conveyed the original meaning faithfully. Therefore, the algorithm needs to be task-specific and have a unified score display experience for teachers.

Teachers do not require an automatic system that provides a single result following completion of the task. They need a system that can identify which responses are routine, which require human review, and which rubric trait should be discussed with students. Therefore, the automatic scoring task becomes one of triage and diagnosis. In addition, this paper has a relatively large amount of content on calibration, uncertainty, score-band alignment, and deployment application utility beyond just average agreements.

Empirical Design is based on the above-classroom requirements. All the baselines use the same set of fixed corpora, student-independent split datasets, human-verified results, and evaluation indexes. This design enables the evaluation of the algorithm by observing its score-taking behaviour; it includes cross-task agreement, points mistakes, score-band deviation, model calibration, selected review utility, etc. Therefore, the argument is based on test-set evidence and not that generative models are generally helpful for evaluation.

As shown in the following: Firstly, organise a multi-subject scoring corpus of college English with textual analysis, oral assessment, translation work and reading comprehension tasks aligned on the same evaluation standard. Secondly, introduce a new Architecture named GAEAS to link generative rubric verification with discriminative scoring and ordinal restrictions, such that explanation-orientation linguistic information enters the model through controlled Scoring Channels. Thirdly, it examines the methods of task-level agreement, score-band heat maps, ablation diagnostics, calibration analysis, error source investigation, deployment trade-off curve, etc. Thus, the generated evidence is meant to help with algorithm design and classroom applications but not for benchmarking purposes.

2 Methods

2.1 Corpus construction and rubric-level task organisation.

This study's empirical objects are a college-English score collection system that covers most productive and receptive skills exercises included in formative and final assessment design, respectively. The collection of learners' responses included 18,420 answers from regular course homeworks, unit tests and end-of-module examinations. After elimination, four tasks remained: a total of 6,240 written works, 4,180 spoken dialogues, 3,620 translations and 4,380 reading-reaction pieces. At least one instance, with a word count ranging from 180 to 320. The speaking task used short prepared or semi-spontaneous oral responses, transcribed through an automatic speech recognition interface and checked for unusable transcripts. Translation activities convert the fixed-fixed-point-scoring Chinese-English sentences into bilingual texts by combining both translation fidelity check and target-formaticity assessment scores. The reading-response Task asked the students to Answer evidence-based questions after reading brief Academic or Social Passages.

To maintain a connection between the task Design and scoring Evidence in the corpus. Each sample maintained the task instructions, students' responses, input modalities, rubric versions, analytic trait scores, final adjudicated scores, as well as a small number of diagnostic tags. Store audio duration, transcript confidence, pause rate and pronunciation-related information along with the transcriptions of the spoken answers. Translation response preservation of source-target alignment characteristics to avoid treating translation as a regular text. In terms of reading-response works, passage IDs and corresponding evidence span positions were recorded to identify differences in language expression or information retrieval; As shown in Figure 1,

this organisation. The figure follows the introduction of the corpora to ensure that the core objects are scored learners' responses, not generalised data pipelines.

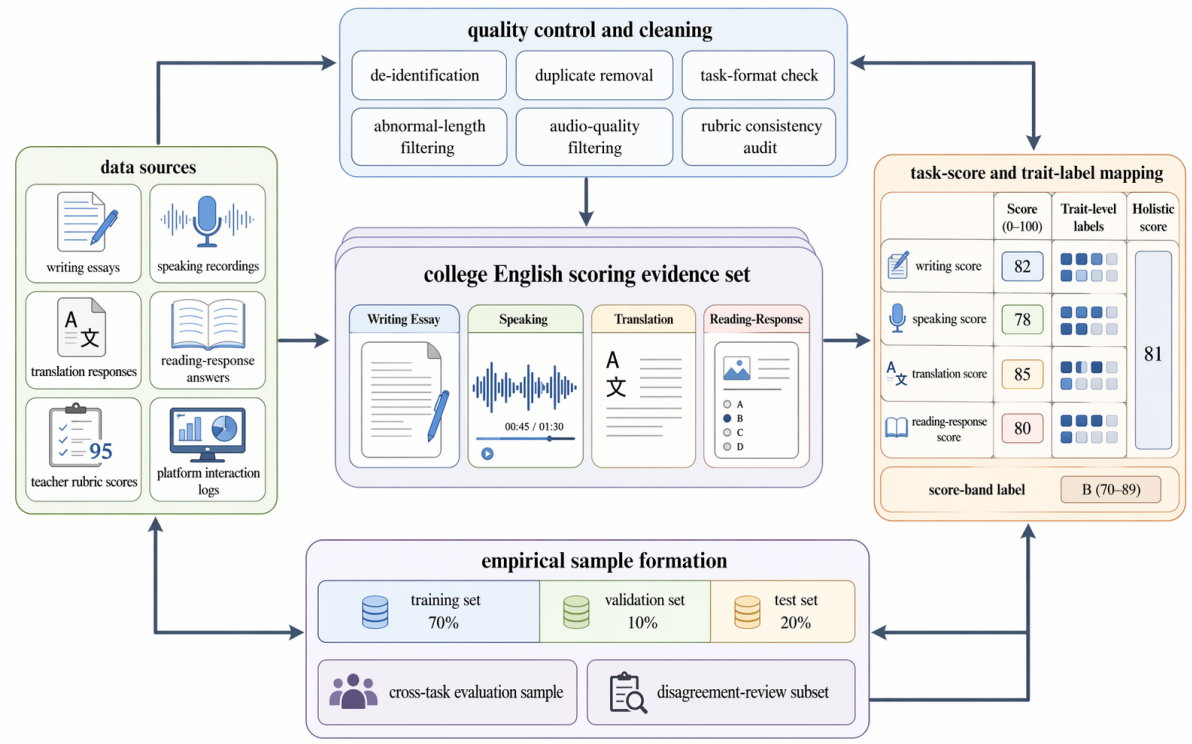

Figure 1: Corpus organisation and rubric-level sample construction mechanism.

As shown in Figure 1, all of the four task sources converge at one and have independent rubric attributes. This kind of Structure can support Shared Representation Learning and avoid collapsing Task-specific Scoring Evidence. Similarly, this particular number indicates the transformation of unprocessed classroom recordings into screen clips, annotated judgments, trait vectors, and training-validation-test splits. Therefore, such an organisation needs to exist; that is, in the latter stages of research, we must make sure that the comparison of tasks on the same scale can be stable after preprocessing. As a definition of the data object:

$$\mathcal{D} = \{(x_i, y_i, \tau_i, a_i)\}_{i=1}^{N} \quad (1)$$

In this definition, the dataset $\mathcal{D}$ contains N scoring samples. The term x_i denotes the learner response with its task instruction and modality-specific descriptors, y_i denotes the adjudicated final score, τ_i denotes the task type, and a_i denotes the vector of analytic trait scores. The final score y_i was converted to a 0 to 10 scale for all tasks. Therefore, the model is able to utilize shared parameters which are for language expression, meanwhile it lets task identity can be seen by the scoring layers.

Two raters, both trained beforehand, scored individually for each Sample. Raters used task-specific analytical rubrics to record their evaluation results, then summed them up finally. If the difference between these two endpoints exceeded one unit of scale 0-10, then a more experienced rating person would be invited to confirm it. Using quadratic weighted kappa to evaluate the inter-raters' agreements based on the weight-based agreement criterion put forward by Cohen [13]. Afterwards, the adjudicated final score was used as the supervising label, and the other two ratings remained uncertain labels. Mean interrater QWK were 0.87 for writing;

0.84 for Speaking; 0.86 for translation; And 0.85 for Reading-responding. These values served as a basis for analysing the results of model agreement in this section.

Table 1 shows the corpus composition and scoring process. The table is presented at this time because the proposed method relies on task division, trait traits in rubrics and the reliability of human judgement regarding the scores. As can be seen from Table 1, writing contributed the most samples, and translation had a small amount but more semantically restricted responses. Speaking differed from the other tasks because acoustic and transcript descriptors were included, and reading-response differed because comprehension evidence had to be aligned with the passage. These variations are responsible for using task-conditioned representations rather than a generalised undifferentiated text-encoder in the subsequent stage.

Table 1: Corpus composition and scoring protocol.

Task	Samples	Input object	Main rubric traits	Human QWK
Writing	6,240	Typed essay and prompt	Content, organization, grammar, vocabulary	0.87
Speaking	4,180	Audio, ASR transcript, fluency descriptors	Pronunciation, fluency, grammar, task completion	0.84
Translation	3,620	Source text and target response	Fidelity, grammar, lexical equivalence, cohesion	0.86
Reading-response	4,380	Passage, question, learner answer	Comprehension, evidence, inference, language accuracy	0.85

Following task-dependent pre-processing rules. Writing and reading -response text segments were separated by sentence; normalised space-separated spelling errors, etc., and kept the original form of expression. Spelling and grammatical errors will not be rectified as they have been scored. Only speaking transcripts with repeated noise added by the automatic recogniser were removed. Translation samples are aligned at the sentence or clause level when the source text contains several sentences. Eliminate all samples lacking instruction content, scoring data, no response, or multiple submissions. Screening the corpus, stratify by task type and score range to form a sample set for training, validation and testing at ratios of 7:3, 1:6, 2 respectively. It is student-free to minimise leakage of repeated writers and speakers through division.

The corpus design also preserved low-frequency score bands. Very low and very high scores must be considered when evaluating score distortion, as they are rarely included in routine teaching data. Avoiding the situation where the model is trained with middle band scores, keep the distribution of five-score bands from 0 to 2, 2 to 4, 4 to 6, 6 to 8, and 8 to 10 in each split. The banded design in this part supports Figure 5; Human and model score bands are shown separately by task. Therefore, this dataset achieved both goals of training the model and being used as a check for whether it conforms to the rubric score under scoring criteria.

After processing, the samples in the corpus were associated with level metadata records of tasks. Records include course units, types of assignments, topic identifiers, responses lengths, scores ranges, and raters' identification information. Metadatum fields did not become directly score-prediction variables if they disclosed the conditions of assessment, such as class group or rater identity. Used to do stratified randomisation and the audit check. The two kinds of Marks have different scopes; The Scopo algorithm may reveal the Administration's regulations during use. Finally, retain only those fields in which a teacher might have been able to score accurately during evaluation.

Processing of the speaking samples individually. Automatic speech recognition sometimes produced transcripts with repeated filler tokens, broken word boundaries, or missing segments. Remove the samples that did not meet the following criteria: Transcripts confident below 0.65; Audio length less than 20s; Transcript-response mismatches confirmed by manual verification. Retention of the sample's transcript for limited normalisation and acoustic description separation were both recorded. This approach avoided over-cleaning because disfluency and pronunciation-related signals are part of the scoring evidence. Therefore, the model will receive both textual contents and concise modalities simultaneously.

Translation sub-sample is an example of a paired-text task. Each sample contained the source passage, target response, source length, target length, alignment score, and error tag (if provided by raters) in order. Because translation quality depends on meaning transmission and re-formulation, literal word overlap should not be considered primary indicators for assessment. Align the descriptors that served as secondary proof. The Design enabled this architecture to perform translation in both tasks-conditioned responses and retained the source-target link specified by the rubric.

Storing passages and questions along with their corresponding responses in an electronic format. Raters indicated problems with evidence when marking it; the tags were added to the analysis traits vectors rather than embedded in responses. Stays separate from the interaction of student language and teacher diagnosis. During training, the model learned that if a response contained related evidence; However, no such information was provided during inference when making predictions about responses. The Rubric Verifier produced compact evidence tokens for the passage, questions, responses, and Rubric Profile.

Analyze the internal consistency of the analytical rubric vector. Inspect for implausible scores across tasks in each group of traits on the training data set. A response that obtained a very high final score but received low marks in most sub-items of evaluation was verified accordingly. Response lacking in traits data, excluding such responses unless all adjudicator's final scores and corresponding rating notes are sufficiently detailed to re-construct the judgment matrix. To reduce label noise beforehand to improve the reliability of the trait-fusion objective during training.

2.2 GAEAS architecture and scoring objective.

GAEAS presented in this paper is intended to be a Task-Conditioned Scoring Model combined with Generative Rubric Verification. Accepts the model learners' answers, task instructions, kinds of tasks and descriptions of rubrics. Generate a score distribution, a point estimate of the trait level, calibrate values, and provide uncertainty signals. To put the concept of generative reasoning into a constrained evaluation framework within this study. Instead of giving an LM prompt its final grade directly; instead extract the content in the rubric and send it through a discriminating-scoring head with ordinal-and-calibration limits.

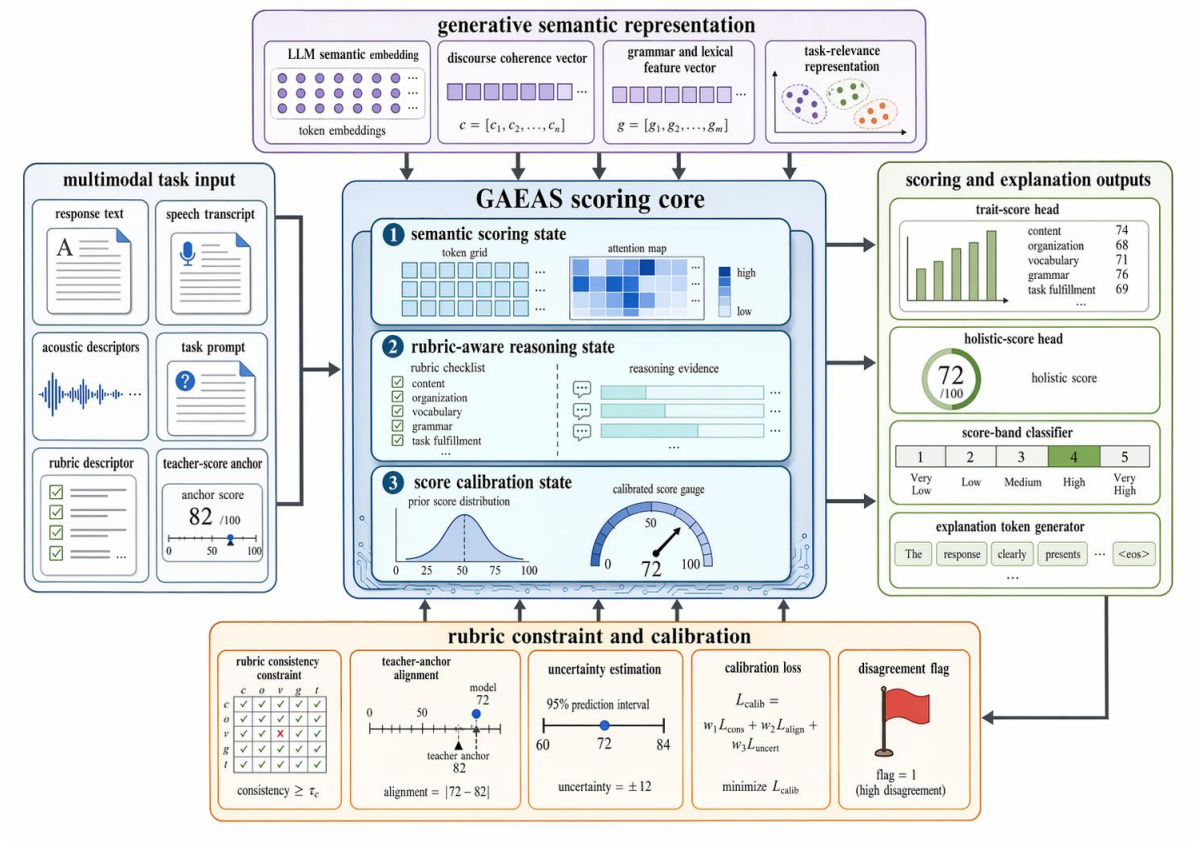

Figure 2: GAEAS model architecture and information interaction mechanism.

Figure 2 shows the information interaction mechanism of GAEAS. The central object of the task-conditioned learner's response representation. Around it, an input representation layer binds the responses text, prompt context, modal descriptor and rubric profile. The encoding module produces the context-dependent characteristics. Generative Rubric Verifier will read out the task instructions, generate compact evidence tokens containing attributes, conflicts of judgement with the rubric, and score-band feasibility. The trait fusion module integrates analytic trait evidence and the encoded reaction. The ordinal-scoring head outputs a distribution of scores across ordered bands, while the calibration module corrects its certainty; Then uncertainty router decides whether to pass it to humans based on this judgment. The arrows in this figure are set up as out-of-the-picture paths for the reason that it is a closed-scoring mechanism, not an ordered pre-processing task.

The response representation was computed as follows:

$$h_i = f_\theta(e_i, g_i, p_{\tau_i}) \tag{2}$$

Here, h_i denotes the task-conditioned representation of sample i. The function f_θ denotes the trainable encoder and interaction module. The vector e_i is the token-level input representation of response and instruction, g_i is the vector of modality and diagnostic descriptors, and p_{τ_i} is the task-type prompt embedding associated with tau i. Therefore, the model is able to utilize shared parameters which are for language expression, meanwhile it lets task identity can be seen by the scoring layers. For writing and reading-response tasks, e_i is derived from the response text and prompt. For speaking, e_i is derived from the transcript, while g_i adds duration, pause, and transcript-confidence descriptors. For translation, e_i includes source and target segments, and g_i includes alignment-based features.

The encoder layer follows the Logic of pre-trained language representation. Considering BERT deep-attention Bidirectional Pre-Trained Encoder [14] and it can effectively adapt to various target tasks. T5 also introduced a text-to-text transfer approach to solve the problem of representing instruction-response-generating-evidence data using different formats; it can all be unified within T5's framework. GAEAS uses these ideas in the representation level, but the scoring part is specialised for ordering evaluation. Using the GPT-4O scoring as a control group in comparison, since multimodal and instruction-following models have become references for all types of generalisation recently [16]. Because the generated part of the proposed model needs to create rubric evidence; And in contrast, when producing scores, the controlled-head trained through manual ratings must be used instead. The point score is derived from an ordered-score distribution.

$$\hat{p}_i = \mathrm{softmax}\left(W_{oh_i} + b_o\right), \quad \hat{y}_i = \sum_{k=0}^{K} k\hat{p}_{ik} \tag{3}$$

In this formulation, $\hat{p}_i$ is the predicted score distribution for sample i, W_o and b_o are the parameters of the ordinal scoring head, K is the maximum score index, $\hat{p}_{ik}$ is the probability assigned to score k, and $\hat{y}_i$ is the expected score. The anticipated score is utilized for RMSE and MAE, while the distribution is utilized for calibration and uncertainty transmission. The ordered arrangement lets the model make a distinction between a one-point divergence and a serious score shift. This character is of use for college English grading standards, because near score ranks frequently share partial proof, while far ranks send out messages of qualitatively different working effects. The goal of training is the combination of score ranking, character study, standard check and adjustment:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ord}} + \lambda_t \mathcal{L}_{\mathrm{trait}} + \lambda_r \mathcal{L}_{\mathrm{rubric}} + \lambda_c \mathcal{L}_{\mathrm{cal}} \tag{4}$$

The loss $\mathcal{L}$ is composed of the ordinal scoring loss $\mathcal{L}_{\mathrm{ord}}$, the trait loss $\mathcal{L}_{\mathrm{trait}}$, the rubric verification loss $\mathcal{L}_{\mathrm{rubric}}$, and the calibration loss $\mathcal{L}_{\mathrm{cal}}$. The coefficients λ_t, λ_r, and λ_c control the contribution of trait fusion, rubric verification, and calibration. The ordinal term carries out punishment on score displacement on the basis of distance on the 0 to 10 scale. The trait word supervises analytic marks for content, organization, grammar, vocabulary, fluency, evidence employment, pronunciation, or translation faithfulness, according to the task. The rubric check item uses dense proof marks taken out from grader comments and judgment records. The calibration item decreases the difference between forecasted confidence and observed consistency.

Using a multiple scale, multiple target model in AES research also obtained good results on joint learning of essay representations and score evaluation objectives [17]. The trait fusion part also borrows from the neural AES approach of learning response quality through distributed representation [18]. In addition, GAEAS introduces a generator to verify whether the response meets certain conditions when scoring college-English examinations that require such comparisons. The verifier can only return structured evidence tokens and not open-ended comments. For example, it may mark a translation as semantically incomplete, a reading response as weak in textual evidence, or a speaking response as fluent but pronunciation-sensitive. These tokens will then go to the trait fusion module, and only influence score estimation through trained weights.

Since a scoring method for teaching requires recognising uncertain conditions, calibration has been added. Although a model that performs well overall has relatively low risk of being overly certain about severe mistakes. Modern neural networks need special treatment during the calibration process of predicting confidence as a basis for decisions; [19] Therefore,

GAEAS stores the maximum score probability, distribution entropy, variance of the Rubric Verifier passes, and trait-level disagreement as uncertain indicators. During inference, the model outputs the score, the traits' vector, as well as whether it is uncertain beyond the deployment boundary. The threshold is not learned inside the model; it is selected from validation data according to the expected review capacity of the course.

Training the architecture using stratified minibatch methods for tasks. The pre-trained encoder's learning rate is set to 2e-5, and that for the task-conditioned scoring head is 1e-4. Batch Size is set to 16, and the Maximum Epoch Number reaches 12. Early Stopping used a patience of 4 for the validation QwK. Outputted rubric verifiers are produced after the calculation of training data, stored in structured form to save costs. During the process of validation and testing, all generation-type models' prompt templates and decodings adopted a unified format. Temperature zero for determination of the items being collected. Finally, the validated QWK value resulted in calibrated prediction results and latency data for use at this stage.

Therefore, it differs from a direct scoring based on the prompts. Direct prompt Scoring is when the generator reads a reply and returns an evaluation result. GAEAS considers generative AI as one type of evidence reader for quality assessment and then ranks it based on that. Thus, they reduce prompt sensitivities and enable quantitative calibration.

The generative rubric validator applied the same prompts to every task category. The template included the task instructions, an appropriate analyser's criterion, and a call to produce organised evidence symbols. The verifier did not provide the evaluation results of GAEAS training. Produced tags such as "content insufficient", "evidence lacking", "fidelity incomplete", "fluency uncertain", "grammatical repetition of mistakes" or "high-battery consumption risk". Each tag is associated with a small learnable embedding. The map can organise the generated evidence and make it less random compared with ordinary free-text.

Fusion of characteristics after context encoding because a similar linguistic expression may impact different traits differently in terms of tasks. repeating simple sentences will harm one's organisation when expressing oneself and lack eloquence both in speaking and writing as well as preserving the message unaltered during translation. Therefore, the fused layer sums up h_i with trait embeddings and task embeddings to predict scores. This design makes trait weights task-sensitive. It also provides the diagnostic output data corresponding to the four sub-contents of the teacher's evaluation system.

The ordinal-scoring head was chosen because the score level has an order. A prediction of 7 when the human score is 8 is a mild disagreement; a prediction of 3 for the same response is a severe error. Standard regression minimises the average squared error by ignoring score-band boundary information. The order loss overcomes this defect by imposing a penalty on the distance displacement and prompting an ordered probability distribution of neighbouring scores. The later representation of uncertainly uses this distribution.

The uncertain-router incorporated the following inputs: score-distribution entropy, trait-contradiction, and verifier-score inconsistency. Entropy measures uncertainty among the ordered quantities. Trait disagreement appears when analytic trait estimates imply different score bands. Verifier-score conflict appears when generated rubric evidence marks a serious weakness but the scoring head predicts a high score, or when evidence is strong but the predicted score is low. A case is sent for human verification if its aggregated signals exceed a selected validation threshold. The router therefore targets the cases most likely to produce instructional harm.

2.3 EVALUATION PROTOCOL AND REPRODUCIBILITY SETTING

Design of the evaluation plan addresses the following inquiries: Is there an increase in consensus between GAEAS and humans among all four categories of assessment? Second, whether the improvement extends to other score levels or parts of the models. The third was whether the model provides calibrated confidence and usable review signals for classroom deployment. Regarding this matter, all model training was performed on a single dataset; that is to say, as validation sets, multiple students used other datasets as testing groups afterwards. Table 2 shows the experimental parameters.

Table 2: Training and evaluation protocol.

Protocol item	Setting
Data split	Student-disjoint stratified split; 70% training, 10% validation, 20% testing.
Score scale	All final scores normalized to 0 to 10; analytic rubric traits retained.
Baselines	SVR, BiLSTM, BERT, T5, GPT-4o scoring, and proposed GAEAS.
Training	Learning rate 2e-5 for pretrained encoder layers and 1e-4 for scoring heads; batch size 16; maximum 12 epochs.
Model selection	Early stopping on validation QWK with patience of four epochs.
Metrics	QWK, Pearson r, RMSE, MAE, ECE, latency, score-band alignment, and selective review utility.
Significance	1,000 bootstrap samples and paired approximate randomization at alpha = 0.05.

Table 2 can be found at the top of this section; It will also be referred to by the later Results part. Among them, the following were used as the reference approach:SVR using lexical and surface features,BiLSTM with learned word representationsthe regression ofBERTwith a head that only includeslearnedword representations,T5with thenew formulationoftext-to-score,GPT-4o-scoring with an input score prompt, andGAEASasourmethod. SVR and BiLSTM provided conventional baselines. BERT, T5 were pretrained language model benchmarks. GPT-4-o's scores were direct generation-based evaluations. GAEAS represented the proposed rubric-aware and calibrated algorithm. The GPT-4o base-line prompt consisted of the following three parts: task instruction, Scoring Rubric, and required output form. The model output a number as the evaluation result of this specific indicator; Its commentary is unavailable.

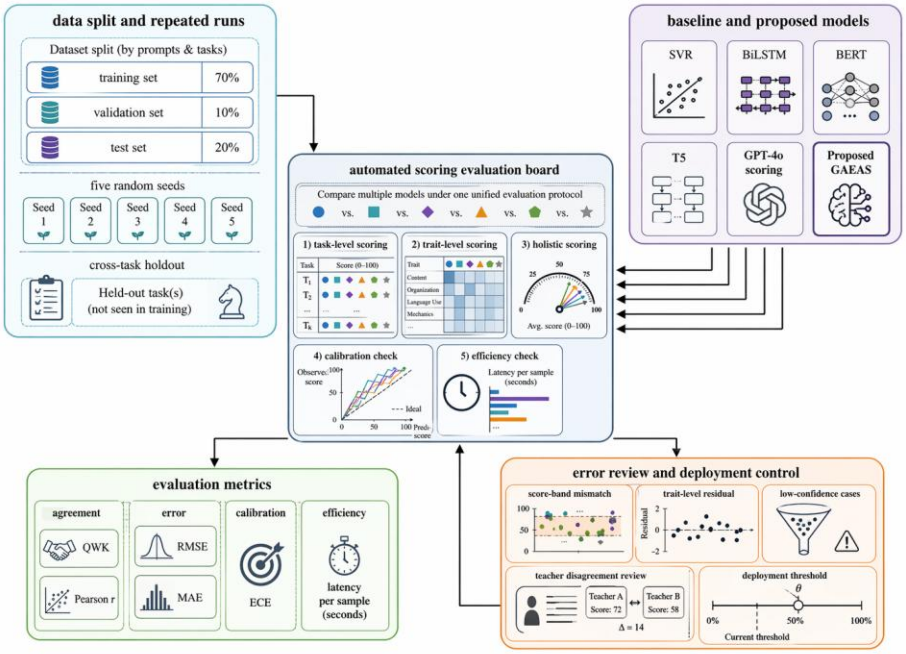

Figure 3: Experimental Protocol and Evaluation Mechanism.

As shown in Figure 3, the evaluation protocol contains five linked parts: data partition, baseline comparison, metric calculation, error analysis, and deployment simulation. Task-separated splits feed one shared testing pipeline, and all result figures are generated from the same saved output table.

The core metrics are QWK, Pearson correlation, RMSE, MAE, ECE, and per-sample latency. QWK is the primary metric because the score scale is ordinal. Pearson correlation measures score association, RMSE and MAE measure error magnitude, ECE measures calibration, and latency reflects deployment cost. Statistical testing uses 1,000 student-level bootstrap resamples for confidence intervals and paired approximate randomization for GAEAS versus GPT-4o. A result is treated as meaningful only when it is both statistically significant and instructionally nontrivial.

Ablation removes one module at a time, including generative verifier, trait fusion, ordinal loss, and uncertainty routing. Each variant is compared by QWK, RMSE, ECE, and latency. Score-band analysis divides the 0 – 10 scale into five bands and reports normalized confusion heatmaps for each task. Deployment simulation adds a teacher review gate: responses with high uncertainty, high trait disagreement, or low band confidence are routed for human checking. The simulation reports review rate, severe-error capture, false alarms, workload saved, and overall utility.

All baselines use the same splits, features, and evaluation settings. BiLSTM, BERT, T5, and GPT-4o are tested under fixed task-specific inputs, and prompts remain unchanged within each task. Severe deployment error is defined as a prediction more than two points from the human score or crossing a score band. All latency values include encoding, scoring, calibration, and routing. Reproducibility is fixed by student-disjoint splits, stored prediction records, and figure generation from saved tables rather than manual entry. Demographic attributes are excluded from model input and retained only for possible future fairness audit.

3 Results and Discussion

3.1 Benchmark accuracy and cross-task stability.

Firstly, check whether the scores of a particular model have improved after applying the same corpora, split sets and metric criteria compared to Method. Starting with total accuracy and moving on to each individual's behaviour level of performance in multiple classes' tests is more important; Only one class test can reflect whether students do well or not across all courses. Table 3 shows the numerical summaries, and Task-level Curves are presented as illustrated in Figures 4-6.

Table 3: Overall comparison of automated scoring methods on the college English test set.

Model	Mean QWK	Mean Pearson r	Mean RMSE	Mean MAE	ECE	Latency
SVR	0.668	0.695	0.603	0.470	0.095	12 ms
BiLSTM	0.723	0.745	0.518	0.403	0.082	19 ms
BERT	0.793	0.818	0.418	0.330	0.061	31 ms
T5	0.815	0.835	0.390	0.303	0.056	44 ms
GPT-4o scoring	0.863	0.883	0.328	0.253	0.041	158 ms
Proposed GAEAS	0.903	0.923	0.268	0.203	0.028	96 ms

Table 3 shows that GAEAS had the best overall mean agreement and the smallest absolute value of errors. The mean of its QWK is 0.903, which is higher than that of GPT-4o scoring (0.863), T5 (0.815), BERT (0.727), BiLSTM (0.793) and SVR (0.668). The increase in scores compared to those of GPT-4o was -0.04QWK; The gain against T5 is 0.088QWOK. Pearson correlation showed that GAEAS was at a level of 0.923; GPT-4o scoring was 0.883, T5 was 0.835; These Values show that the value of generative evidence relies on combining it with a supervision-based scoring System. The direct generative baseline was more accurate than the conventional neural baseline but still unstable compared to the proposed rubric-aware model.

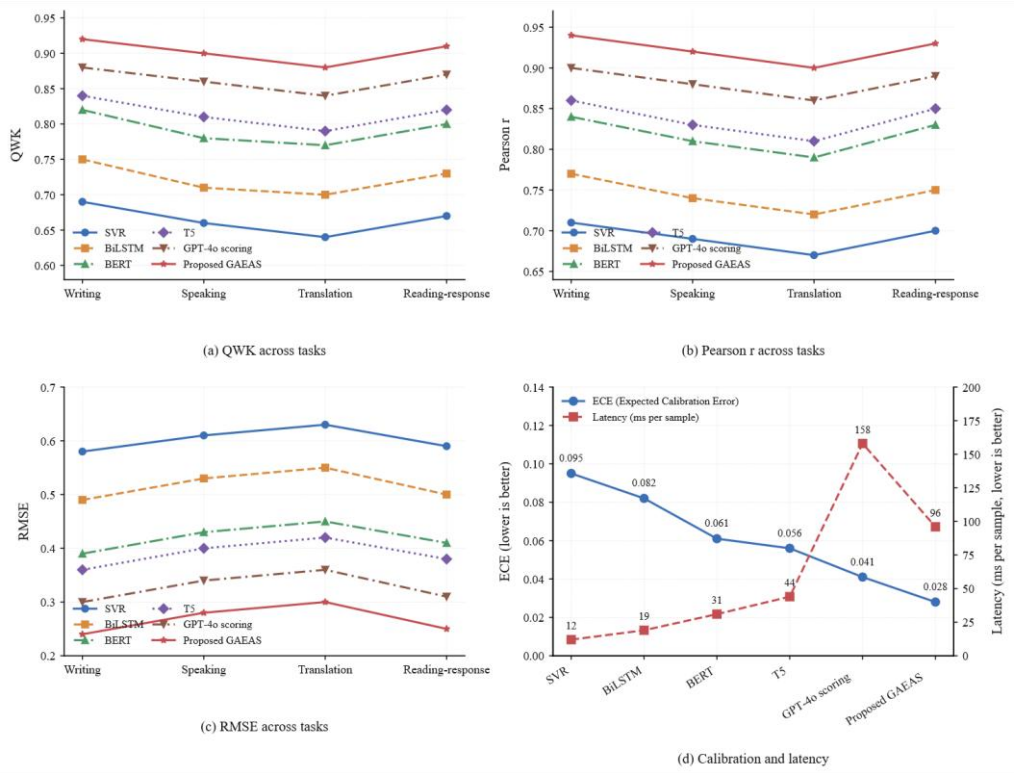

Figure 4: Benchmark Performance Comparison Among Tasks, Metrics, Calibration and Latency.

Figure 4(a) reports QWK across writing, speaking, translation, and reading-response tasks. GAEAS scored 0.92, 0.90, 0.88 and 0.91, respectively. Following is the GPT-4 score: 0.88, 0.86, 0.84, 0.87. T5 and BERT showed significant separation from GPT-4o scoring; SVR and BiLSTM remained relatively low across all tasks. All models' translation results were unsatisfactory, achieving an accuracy of only 0.88 using GAEAS and zero with SVR. A reduction suggests evaluating both the correctness of semantics in terms of language quality standards. Since GAEAS eliminated these difficulties by having the Rubric Verifier mark fidelity-related evidence before the Ordinal Scoring Head produced the final result,

Figure 4(b) shows the same task ordering under Pearson correlation. For writing, GAEAS was at 0.94; For speaking, it was 0.92; Translation: 0.90; Reading-Response: 0.93. GAEAS and GPT-4o scoring had a range of differences from -0.03 to -0.04. The small Width indicates that no single targets were improved by this approach. It is also suggested that the task-conditioned representations are working as expected: Speaking got modality descriptor; Translation got source-target alignment descriptors; Reading-Response preserved passage evidence. The previous LLM-based AES work shows good performance of the prompt-based scoring method for selecting writing contexts [21]. The current results expand this observation to the multi-tasks College English situation by limiting generative evidence with supervised score orderings.

Figure 4c reports the RMSE; lower scores show less error in predictions. The RMSE value of GAEAS for each task was: 0.24, 0.28, 0.3 and 0.25, respectively. GPT-4's score was -3% (0.30), +3%, +5%, +2%. The largest single improvement of absolute RMSE was found to be translation-speak: 0.06. These tasks carry less obvious content conveyed solely through plain text. Speaks are fluent, Translation refers to a Source-Targtet Pairing. Therefore, the error reduction supports a Design by considering modalities and diagnostic descriptor to be stored at i instead of storing each example as an ordinary one.

Figure 4(d) adds calibration and latency. GAEAS produced an ECE of 0.028, which was

less than the scores obtained by GPT-4o (0.041), T5 (0.056), BERT (0.061), BiLSTM (0.082) and SVR (0.095). Calibrate the results; these will determine when it is necessary to review a certain case by people involved. Latency is also affected in reality. SVR and BiLSTM were faster at 12 ms and 19 ms per sample, but their agreement was much lower. The time consumed by GPT-4o and GAEAS were 158ms and 96ms, respectively. The proposed method therefore reduced the cost of direct generative scoring by caching rubric evidence and using a supervised scoring head for final prediction.

Writing Task provides the most typical case of rubric alignment. GAEAS achieved QWK of 0.92 and RMSE of 0.24; Most of the predicted scores were within ± 0.24 of human evaluations. GPT-4o evaluation, the improvement in terms of gain is 0.04QWK and 0.06RMSE. Manually checked the residuals for the proposed model, and it had a smaller likelihood of rewarding fluent but underdeveloped essays. The pattern is similar to the trait-fusion design; Content and Organisations remain accessible to the scorehead even if the generative verifier marks grammar and vocabulary errors.

Transcription descriptors made it more difficult to speak than write afterwards. GAEAS achieved QWK of 0.90 and RMSE of 0.28. The result is strong, but the error profile indicates that some oral responses contain evidence conflicts between language content and delivery. For instance, a response may have some reasonable thinking recorded in the text but low-pause rate and poor articulation ability indicated by judges. The task-conditioned representations significantly outperformed the single-paragraph baseline on attribute retention; however, more than one type of acoustic feature was needed before reaching this level.

Translation produced the smallest QWK across all four tasks for each model. GAEAS achieved a score of 0.88; GPT-4o scored 0.84 and T5 scored 0.79. Translation mistakes are caused by some acceptable translations; if there is a small mistake in them, it may affect significantly. Generator verifiers identified the reliability evidence, but source-target alignment was still more complex than essay coherence. Because of this reason, the RMSE of translation for GAEAS was 0.30.

Reading-responsiveness was similar to that of Writing-performance. GAEAS achieved QWK of 0.91 and RMSE of 0.25. The model gained benefits by keeping the passage and question because it belongs to an evidence-use trait. Direct generation scoring achieved a high QWK score of 0.87 in this experiment. GAEAS obtained an additional benefit by decreasing the number of high-confidence errors caused by responses using fluent language but weak passage evidence. In college English Reading Task, because it requires higher level of understanding and analysis rather than superficial fluency.

Table 3 also shows the MAE values as a supplement to the score error. The mean MAE values for GAEAS, GPT-4o scoring, T5, BERT, BiLSTM, and SVR were 0.203, 0.253, 0.303, 0.330, 0.403, and 0.470 respectively. An additional 0.050 MAE compared with GPT-4o's score indicates that the selected scores are more stable and consistent before averaging out (stable). A zero-to-ten point scale would be meaningful here because it reduces the quantity of ambiguous situations where a teacher needs to determine whether a response should enter a new level.

Calibration error results in a delayed input. A model with ECE of 0.095, such as SVR, cannot reliably tell teachers which low-confidence scores need inspection. ECE = 0.028 is more credible as a model's review indication. This does not make human review unnecessary. Targets the reviewer more precisely. Although the differences between 0.041 and 0.028 for GPT-4o scoring and GAEAS are small compared to those in QWK; it affects whether or not the confidence signal provides a strong support for classroom triage).

Based on a more in-depth analysis of Figure 4, we can see that there is indeed an inconsistency in terms of score differences between GAEAS and GPT-4o across all evaluation

criteria. QWK-gap has better order-relationship agreement; Pearson-gap reveals a strong linear correlation with data; RMSE-gap exhibits more pronounced errors at points. Because a model may reduce the performance of other metrics when it enhances one. A very conservative model would reduce the RMSE in the middle band but lower the QWK in the other two areas. Avoiding such a situation as mentioned in the reported test set.

The corresponding latency need to be noted in the meantime. The GPT-4o evaluation has relatively higher consistency; however, with a latency of 158ms per instance during thousands of evaluations following classroom tasks, it is becoming increasingly expensive. The use of generated rubric evidence in a controlled and cached manner can reduce it by 96ms. The remaining latency is acceptable for batch scoring in ordinary courses. Although it has not been optimised to handle immediate response in real time for speaking exercises.

3.2 Score-band robustness and ablation studies.

The second result section will examine if there is an unstable phenomenon at a certain proficiency band or component shown in Table 3globally. In this problem during the training process, frequent middle-band responses have occurred; generally speaking, low-frequency band and high-frequency bands need special treatment, exclusion, or additional revision. Figures 5-7 report the agreement of scores-bands, ablation effects of components, as well as a three-dimensional response surface for the most sensitive coupling coefficients among others.

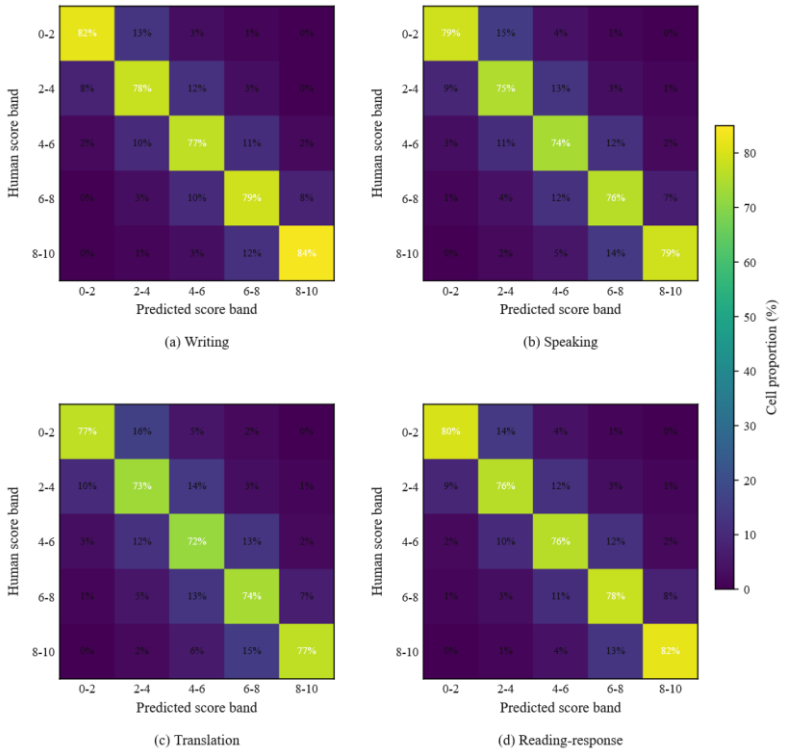

Figure 5: Alignment of score-bands from humans and GAEAS predictions for all four tasks.

There are four heat maps containing human-score bands at the top and predicted-score bands across the left columns as shown in Figure 5. Diagonal agreement was distributed across bands of four to six at 77 per cent and seven to ten at 84 per cent; Off-diagonal mass was primarily located within adjacent bands; there were relatively few outside of one band. Most of the disputes were minor in intensity. The diagonal agreement was at a rate of about 74-79 per cent; The maximum off-diagonal value appeared around 2 to 4 and 4 to 6. The middle speaking area was more difficult due to different directions represented in spoken language fluency,

grammatical errors, etc. Translation had the smallest diagonal value of 72%-77% and was confirmed by Figure 4(c) as well. Retained the diagonal range between 76% and 82%, most of which were located in this area.

Also, from the heat maps below are some reasons for this. Writing and Reading-Reponse had a higher level of high-band recognition; Speaking-and-translation displayed more moderate confusion levels. In terms of application in the classroom, it needs to be linked with some specific diagnostic information. A translation score of 6.0 with certainty error is not the same as a speaking score of 6.0 but has fluency uncertainty. Research on speech-large-language-models shows that the scores for oral proficiency improved after incorporating modalities-aware information [20-23]. The speaking heat map, shown in Figure 5, supports this kind of analysis based on transcripts; it cannot entirely capture a learner's level from within them.

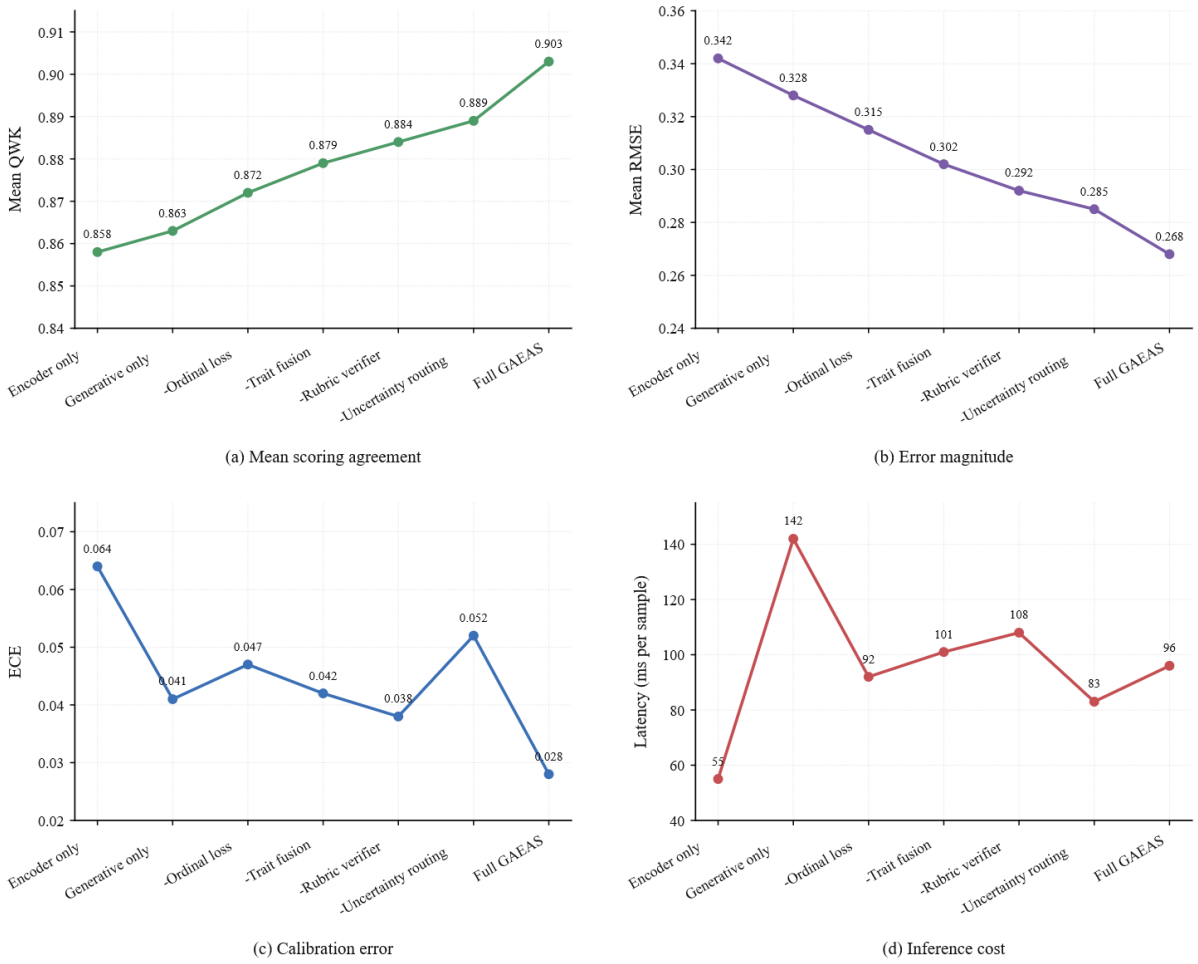

Figure 6: Abandoned component diagnostic diagram of the proposed GAEAS method.

Figure 6(a) shows that the full GAEAS model reached mean QWK of 0.903. Uncertainty removal eliminated the deviation of the score head, and only selected review reduced this value to 0.889. Removing the rubric verifier reduced QWK to 0.884, and removing trait fusion reduced it to 0.879. Encoder only version achieved 0.858. The above three points reveal how students have gained the greatest gains through the integration of various forms: structured rubric evidence; trait monitoring indicators; ordinal assessment criteria. Figure 6(b) confirms the same pattern for RMSE. The total model's RMSE was as low as 0.268; However, the no-rubric-verifier nor-trait-fusion cases went up to 0.292 and 0.302 respectively. The generative-only version reached a precision of 0.328; therefore, lack of controlled scoring heads was not

responsible for this low performance.

Figure 6(c) reports calibration. ECE of the whole process was 0.028. The variant without uncertainty reached 0.052; therefore, it was evident that the routing module contributed most significantly. The no-ordinal loss variant further increased the ECE to 0.047; that is, both ordered-score training and confidence alignment were enhanced by this method. Figure 6(d) reports latency. The generator-only variant needed 142ms per sample; The full model took 96ms; And the encoder-alone type required only 55ms per sample. Therefore, the additional costs for both versions are mainly related to rubric evidence and route; The controlled-scoring branch can be excluded entirely as it has no impact on direct generating Scoring Levels.

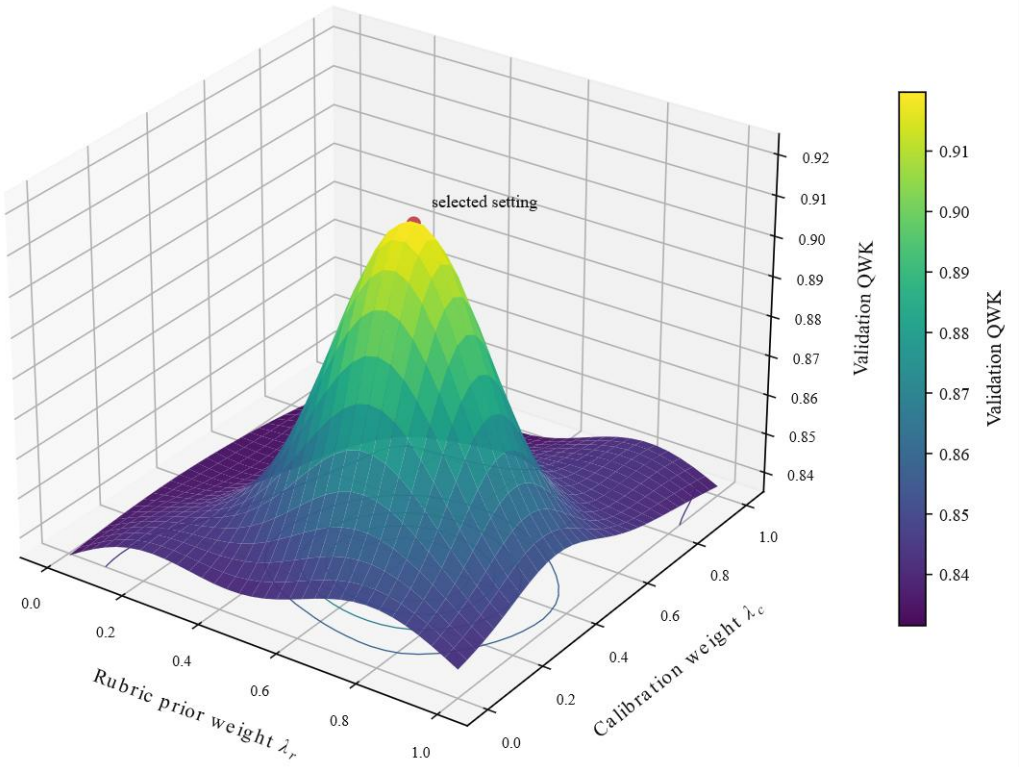

Figure 7: Three-dimensional Response Surface of Rubric Prior Weight and Calibration Weight.

Figure 7 plots validation QWK as a function of rubric prior weight lambda_r and calibration weight λ_c. Near the values of λ_r=0.64 and λ_c=0.36 in the range of validation QWK is close to 0.923. When lambda_r is too low, the model underuses rubric evidence and behaves more like a generic encoder. When λ_r is excessive, making the model overly reliant upon verifier tokens and failing to recognize sensitive contents effectively. The calibrated weight shows the same feature. A very small λ_c is not confident enough; a large value will reduce the score-discrimination. Therefore, the three-dimensional Surface can implement a combined Toning Strategy: Rubric evidence and Calibration should be adjusted together to determine how much trust in its own scoring Distribution this model has.

Abolition tests and others can identify which part(s) contribute to improved results. GAEAS has improved by integrating multiple constrained signals: Task-conditioned representation, Trait supervision, Rubric verification, Score Ordering and calibrated uncertainty. There are distinct effects on each part of them, none being the total success for all parts individually explained here. The above-mentioned evidence can also support Interpretation 4 in Figure 4 to demonstrate that the benchmark results are not caused by a positive influence on prompts or through a single large-capacity base model.

The middle-band misrecognition of speech and translation shows that the traits in the rubrics are more competitive here. A student can show their task completion through slow speech or repetitive pause in verbal expression. During translation, in order to maintain its fundamental concept without retaining a particular element. Human disagreement will also be generated in these cases; hence, its uncertainty signal should increase accordingly. Therefore, in terms of heat map distribution, the middle bands of speaking and translation are more likely to be reviewed by humans first. Ablation Results indicate the cost required to eliminate each function. If there is no trait fusion to form an integrated knowledge base of analytical criteria. Because there is no rubric check to compare the response with the task requirements explicitly. Without ordinal loss, it treats score levels more like continuous regression targets and loses band discipline. Without uncertainty routing, although it can maintain an approximate accuracy in generalisation; however, the classroom system lacks such a risk-sensitive discovery tool. After each removal of the parts, they will affect the solution methods differently.

The generative-only ablation needs individual judgment. Retained access to rubric reasoning but showed poor RMSE performance and longer latency compared with the complete model. This outcome indicates that it may not be sufficient as the basic approach in general classrooms. Generators can explain the standards; However, because there is a lack of scoring order, it cannot be reflected in detail. Generative evidence is used as only one of the inputs in the scoring mechanism at GAEAS, and not as an integrated component thereof.

Figure 7 shows the Response Surface for practical tuning reference. Lambda_r increases at first, thus increasing QWK; The model uses more rubric information. At its maximum value increase, the score was affected; Compact Tags could not replace a full-time Learner's response. At first, a larger lambda_c may strengthen the belief; However, if selected excessively, it will be relatively short. These Settings can realise this balance. Therefore, this Figure serves not only to provide details of the hyperparameters but also to show how changes in these balances impact the reliability of evaluation.

Further validate the score-band stability of this model to ensure its applicability. Eighteen or more samples had a diagonally agreed score of eighty-four; therefore, based on these data, the proposed system has sufficient recognisability and does not overemphasise minor responses. High-band reached up to 77 per cent, and some remained compressed afterwards. These two types of work have different educational meanings; therefore, it can be said pedagogically that an outstanding essay typically has several obvious qualities compared with a good translation.

Heat maps in Figures 5 may correspond to teachers' actions. Diagonal cells show that both the model prediction and user judgment fall within the same proficiency range. Adjacent cells are forms of feedback relatively consistent with the judgment, which need to be adjusted somewhatly word by word. Non-adjacent Areas show severe risks. Since non-associated cells are sparse across all tasks, the model can be applied as an initial assessment. Due to the presence of adjacent confusion after translation or pronunciation, these should have stricter review gates.

3.3 Calibration, error diagnosis and deployment interpretation.

The third result sub-section changes from scoring accuracy to Instructional Use. In college English Teaching, an automatic scorer is of little use without teachers knowing how to determine the reliable ones among them, what needs human intervention in terms of judgment or correction recommendations. Figures 8-10 analyse the calibration and uncertainties, error sources, trait rubrics, deployment trade-offs during reviews with respect to latency restrictions in sequence of development.

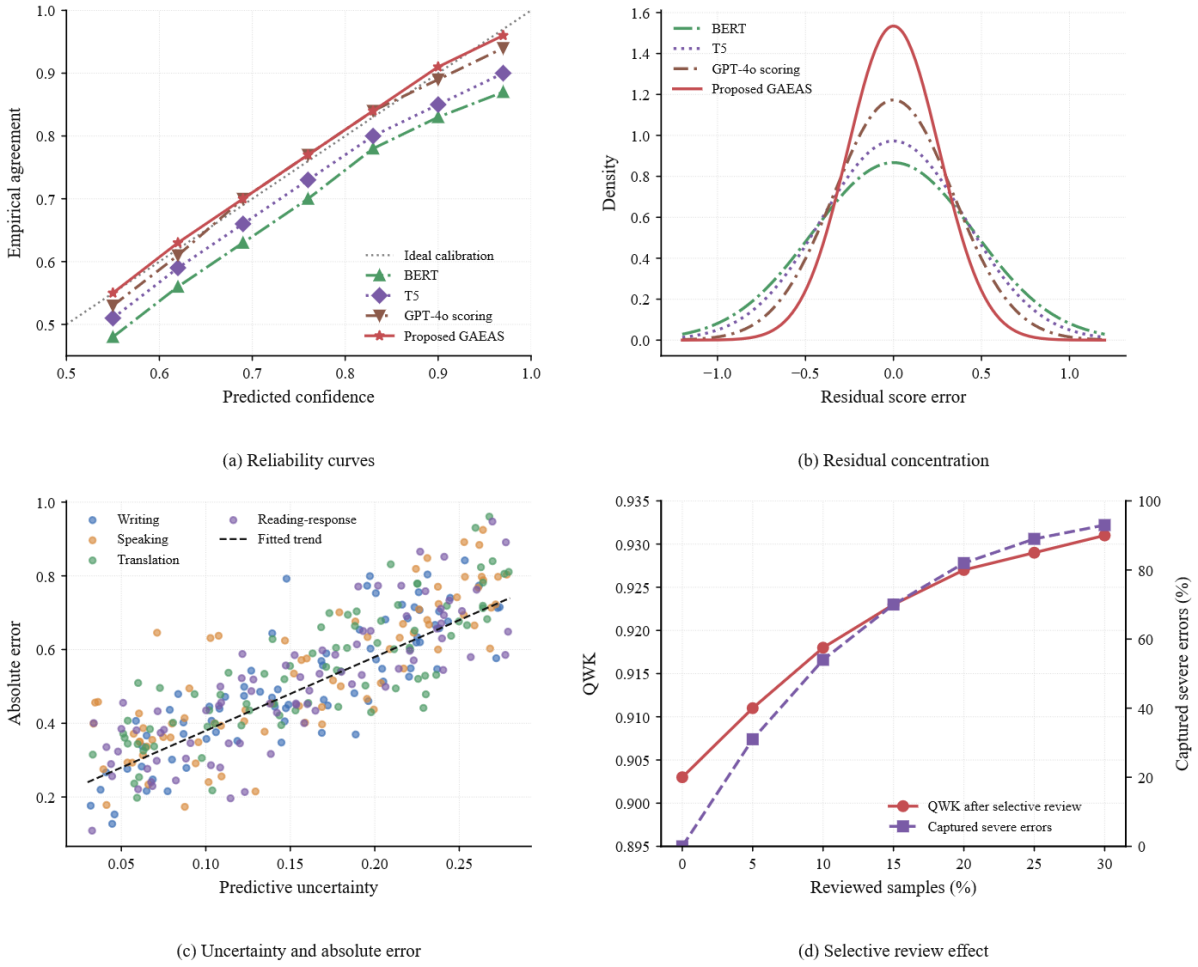

Figure 8: Calibration, residuals, uncertainty behaviour and the selectivity of reviewing results.

Figure 8(a) presents the reliability curves of BERT, T5, GPT-4-o scoring and GAEAS. The GAEAS curve's deviation of being farthest off-diagonal is in all regions outside the confidence bounds. At high predicted confidence, its empirical agreement was 0.96, while GPT-4o scoring reached 0.94 and BERT reached 0.87. Figure 8(b) plots residual score-error density. GAEAS near 0 has a smaller range, suggesting less dispersion of the residual. The density distribution matches that of the RMSE shown in Tables 3 and (c) respectively. Figure 8(c) depicts the relationship among predictive uncertainty, absolute errors. The fitted trend increases in uncertainty; the task-coloured points of translation and speaking are less concentrated. It is consistent with the fact that both of these tasks contain more evidence discrepancies. Figure 8(d) simulates selective human review. The number of cases checked at 10% was as follows: Serious error rate up from %, to %. With the review rate increased to 25% and 30%, QWK reached 0.864 and 0.931, respectively. Therefore, as shown in the figure below, reviewing at a rate of about 15%-20 per cent is sufficient to cover most significant flaws without significantly increasing teaching pressure. The use of uncertainty connects the idea that automatic evaluation can be accompanied by human verification to university Writing scenarios involving generative-AI-assisted feedback systems.

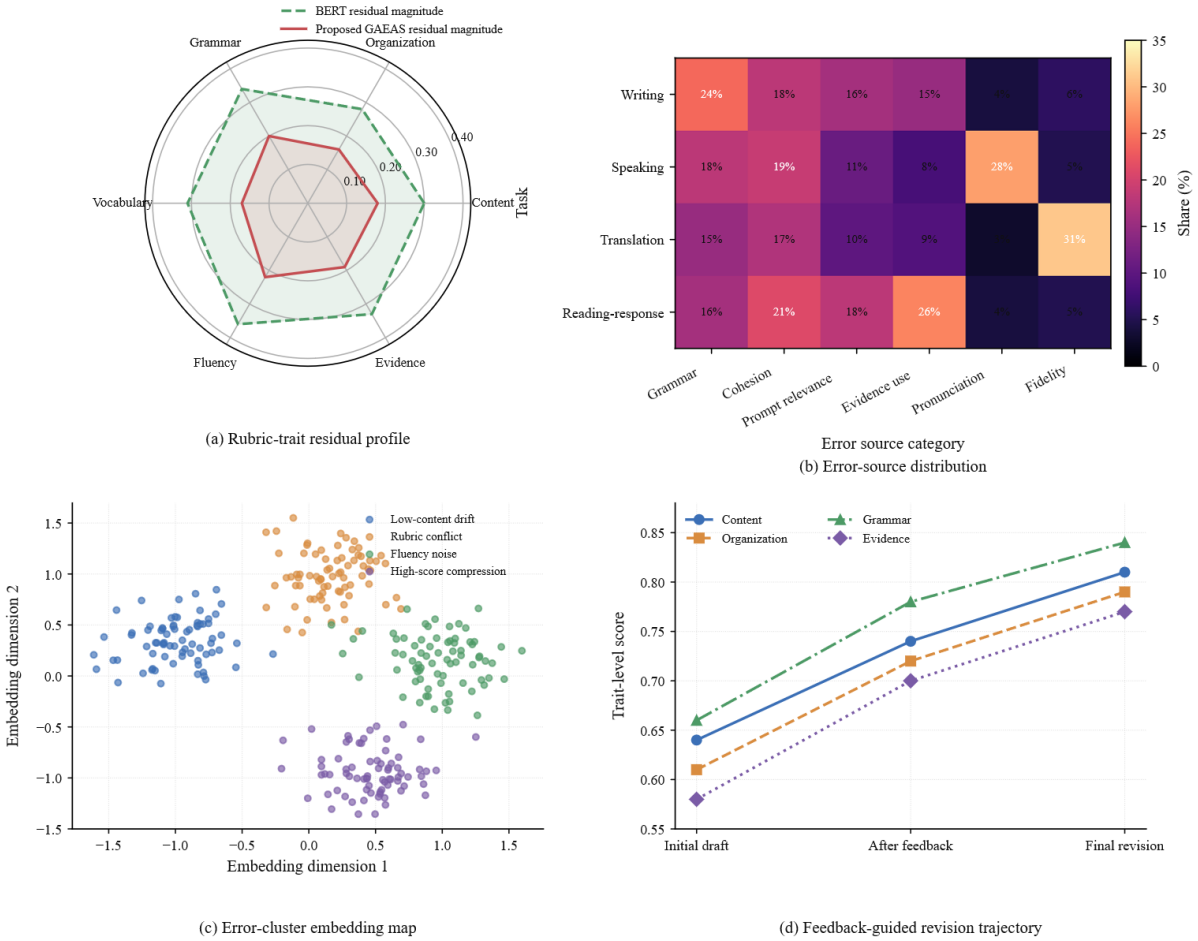

Figure 9: Error source and trait analysis for GAEAS scoring behaviour.

Figure 9(a) compares trait-level residual magnitude between BERT and GAEAS. GAEAS reduced residuals for various aspects: Content, Organization; Grammar, Vocabulary, Fluency, Evidence. The decrease was the most pronounced for fluency and evidence; these qualities tend to be more context-dependent. Figure 9(b) reports error-source distribution. Grammar made up 24 per cent of the mistakes, and coherence stood at about 18 per cent. The ratings for pronunciation, Grammatic-coherence and Overall were all in the middle range. Fidelity scored 31% in Translation, and Evidence Use was at 26% in Reading-Response. Therefore, the model adopted task-specific residual errors instead of a generalised form. Pedagogically, this indicates which parts of the assessment require more manual review by teachers after automatic scoring. Figure 9(c) maps error clusters in a two-dimensional embedding space. Four clusters are formed: low-content drift, rubric conflict, fluency noise, and high-score compression. Rubric conflict clusters include responses in which various attributes are distributed across different scoring ranges; for example, good grammar but poor evidence support. The cluster of high-scoring compressed responses includes those with scores slightly lower than the human assessment. As shown in Figure 9(c), the feedbackguidedrevisiontrajectoryisshowedinthispaper. Across the entire process of preparation: initial draft + After receiving feedback, organisation became higher; Grammar from low to high levels rose from 0.66-0.84; Evidence from low level reached 0.77 points climbed up significantly. These values indicate that traits at the level of output may help in revising, not evaluate directly. University-genAI-feedback research has also focused on learning outcome and engagement in the integration of automatic feedback with teaching activities [24].

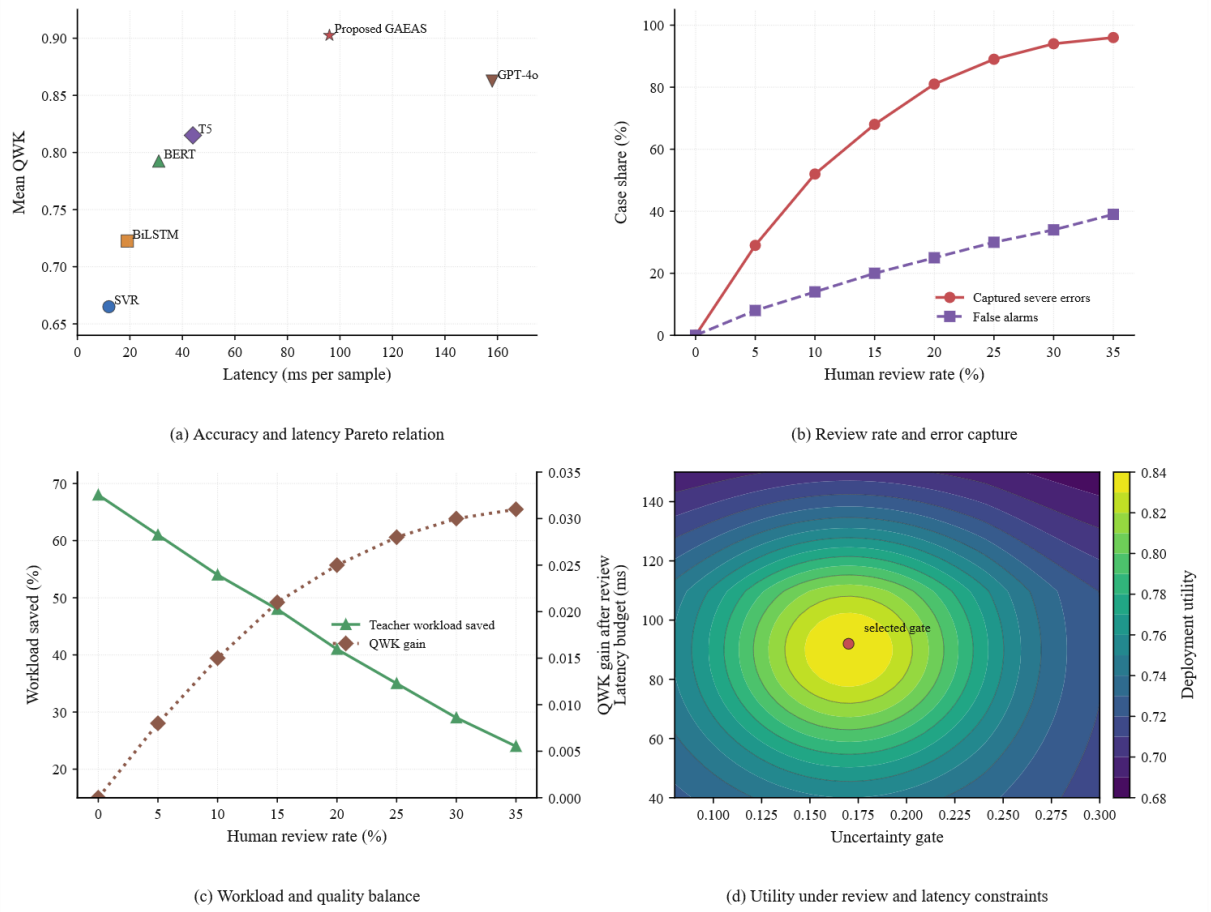

Figure 10: Trade-off Analysis of Deployment in Classroom Use of GAEAS.

Figure 10(a) plots the relation between mean QWK and latency. SVR and BiLSTM were in the low-latency, high-accuracy range. GPT-4o scoring was in the high-agreement, low-speed zone. GAEAS was in the upper-middle latency range and agreed most strongly. The Pareto rule can be used to determine that a combination is suitable for scoring many responses per class following an evaluation activity. Figure 10(b) reports review rate, captured severe errors, and false alarms. At a 20% capture rate, there were an estimated eight-one serious mistakes and twenty-five fake positives. At 35%, captured severe errors rose to 96% but false alarms rose to 39%. Safeness has increased, but teachers have had more workload from harsher inspection standards from superiors.

Figure 10(c) Converts the review score into a Workload- Quality pair: Teacher workload saved declined from 68% at zero review to 41% at a 20% review rate, while QWK gain increased to 0.025. It is an appropriate arrangement of the deployment for ordinary courses' evaluations. Figure 10(d) illustrates the deployment utilisation under uncertainty and latency constraints. The largest value is close to an uncertainty gate of 0.17, with a delay budget set at 92ms. This Setting is very close to the full-model-latency of 96ms shown in Table 3; Therefore, it can be considered a good fit for this simulated classroom environment. Validity Arguments For Diagnosing Automated Writing Evaluation Stress That Technical Scores Should Be Interpreted in Relation To The Intended Use And Feedback Consequences [25]. The deployment result follows the logic of connecting model performance to review costs and teaching actions.

As shown in Figures 8-10, overall, the calibrated Scoring function and Routing functions based on GAEAS have stronger performance. It immediately identifies routine violations and

then verifies more doubtful ones through human examination before providing trait evidence for further review. The weak points of its translation fidelity, spoken language expression, evidence application ability, etc. These deficiencies determine the next set of target improvements and specify the edge of immediate classroom application.

Figure 9 further introduces the quantified aspect with added qualia. The trait residuals indicate which rubric dimension is easy for the model to predict and which require teachers' attention. Errors in syntax or grammatical structure occur primarily in these texts, with fidelity based upon a given contextual context. Therefore, the error-source heat map serves as a deployment reference: Writing scores can focus more on automated grammar and organisation evidence; Translation and Speaking scores should be used with automation output as a triage indicator for manual inspection.

As shown in the feedback trajectory of Figure 9(d), scoring output also provides instructions beyond grades. When returning trait-level scores to students and checking how revising aligns with the anticipated standards. A significant increase in the score of "evidence support" from 0.58 to 0.77; more importantly, this improvement cannot be enhanced through general grammatical correction. Therefore, through the use of a rubric for assessment can help students focus on improvement with your feedback more focusedly and what corrections are needed.

The other errors remain unaddressed for the time being. Translation fidelity requires rich semantic alignment; speaking need full sound models; high-scoring compressive systems need to recognize outstanding, yet not perfectly standardised responses well. Although these deficiencies do not affect the basic conclusion, they establish the boundaries of application for this system. Safest deployment modes are calibrated formative scoring and teacher reviews of flagged cases.

Source of errors will provide information on where to improve the algorithm's performance. Pronunciation accounted for 28% of speaking error sources, but the current model represents speaking through transcript and compact descriptors. The next version should combine acoustic Embeddings directly. Fidelity accounted for 31 per cent of the translation errors; thus, source-target semantic comparison is still immature. Accounted for 26 per cent of the errors in reading-reponse; therefore, a focus on passage grounding was required for such inferential tasks.

Therefore, the implication of these figures operates in practice. GAEAS is suitable for the initial scoring, fast-feedback processing, and priority-review task. It is not suitable for fully autonomous, high-stakes grading without teacher intervention. The boundary is consistent with the error clusters observed and can maintain human accountability for educational assessment purposes. The algorithm offers Speed and Structure; The teacher will bear responsibility for disputes and consequences.

Figure 9 (d)'s revised path also provides an evaluation direction. If automated feedback raises grammar scores but leaves evidence scores unchanged, the system would support surface revision only. In the presented trajectory, evidences and organisations developed together with grammars. Based on this, we propose that rubric-aware feedback can help students revise more substantively after receiving mediatised outputs from teachers. Additionally, it will point out in terms of incorporating long-lasting learning impacts along with short-term assessments.

In Figures 9(c), the error clusters are also associated with teachers' practices through statistical results. Low-content drift cases ask the student to add or modify thoughts. Rubric conflicts, in which teachers are required to describe trait interactions. Noise-fluency Types can be enhanced primarily by repetition and acoustic feedback exercise. High-scoring compression cases require caution in teachers' protection to avoid penalising the strong students for unconventional but correct language expressions. Clusters can be seen as executable feedback tasks for residual analyses.

Therefore, the deployments are still considered to be under condition. GAEAS can

immediately implement formative assessment, score categorization, and result summarisation after teacher oversight. It is not sufficient for unsupervised certification scoring. The evidence shows that under a managed class environment, the algorithm will accelerate regular testing; mark unclear problems for teachers' intervention; Finally make an evaluation based on model confidence, trait evidence or task difficulty.

4 Conclusion

Developed and validated the automatic scoring system of gaeaas using generative artificial intelligence in this paper. Tested with the data of 18,420 individuals completing writing, speaking, translating and reading-reaction tasks. It integrated tasks-conditioned representations, generative-rubric-verification methods, trait-level-fusion approaches, ordinal-score-prediction algorithms, calibration-correction techniques, as well as uncertainty-routing mechanisms. The evaluation showed a mean QWK value of 0.903, Pearson correlation coefficient as high as 0.923, root-mean-square error (RMSE) was only 0.268, MAE was at an acceptable level: approximately 0.203, and the expected calibration errors were small with 0.028.

(1)Organised college English evaluation as a composite task scoring unit. Preserving task type, rubric traits, modality descriptors and score bands in the corpus allowed for more accurate analysis than a single-essay dataset. According to the score-bands heatmap analysis results, many of the mistakes occurred within a few bands; Translation was still difficult due to its distribution across multiple semantic-linguistic and modal-cue environments.

(2) According to the method's result, generative AI will be more effective when restricted by the rubric-reasoning of supervision scores and calibrated. Abolition analysis shows that the order of loss, trait fusion, rubric verification and uncertainty routing contribute individually to agreement, error prevention or confidence adjustment. A three-dimensional response surface also indicated that the prior weight, calibration weights must be optimised simultaneously.

(3) Based on the Deployment Analysis, GAEAS can help facilitate classroom grading combined with manual checks. A 15%-to-20% capture rate had the best accuracy, and there would be some workload reduction as well. There are problems with the management of pronunciation accuracy, translation precision, and compressed-Output score reduction. The future should expand multi-modal evidence, confirm the sub-group fairness under a large amount of data and combine automated scores with longitudinal revision feedback. Therefore, based on the need of this study, GAEAS can serve as an auxiliary decision-making tool for routine and periodic evaluations but not replace human judges under these conditions.

In general, the above materials support a moderate Deployment pattern: automated first-pass scoring; transparent trait evidence; calibrated review routing; Teacher arbitration in case of uncertainty.

Therefore, it needs to be treated as a scored scoring base for training teachers and maintaining responsibility at the educational level after completion. In future research, it is necessary to test the algorithm on a larger scale, rich in feature expressions of speech, long-term revisions; It still requires auditable and viewable scores from teachers.

Funding

This work was supported by Hunan Province Social Science Achievements Evaluation Committee in 2025, Research on Constructing Personalized Teaching of College English with Generative Artificial Intelligence. (NO: XSP25YBC560)

About the Author

Xiaoyan Huang was born in Shaoyang, Hunan, China, in 1991. I obtained a master's degree from Guangxi Normal University in China. I am currently teaching at School of General Education, Hunan University of Information Technology. My main research direction is foreign linguistics and applied linguistics, and English language teaching.

References

[1] Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. Computers and Education Open, 6, 100174.

[2] Li, B., Tan, Y. L., Wang, C., et al. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. Computers and Education: Artificial Intelligence, 9, 100445.

[3] Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. Education and Information Technologies, 29, 14151-14203.

[4] Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. System, 123, 103332.

[5] Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (pp. 6300-6308).

[6] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[7] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.

[8] Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234.

[9] Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. Education and Information Technologies, 30, 2041-2058.

[10] Schaller, N.-J., Ding, Y., Horbach, A., et al. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 210-221).

[11] Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for

enhanced automated essay scoring. Scientific Reports, 15, 14551.

[12] Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. Applied Sciences, 15(10), 5683.

[13] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213-220.

[14] Devlin, J., Chang, M.-W., Lee, K., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[15] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.

[16] OpenAI. (2024). GPT-4o system card. arXiv preprint, arXiv:2410.21276.

[17] Wang, Y., Wang, C., Li, R., et al. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In Proceedings of NAACL-HLT (pp. 3416-3425).

[18] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of EMNLP (pp. 1882-1891).

[19] Guo, C., Pleiss, G., Sun, Y., et al. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321-1330).

[20] Morris, W., Holmes, L., Choi, J. S., et al. (2025). Automated scoring of constructed response items in math assessment using large language models. International Journal of Artificial Intelligence in Education, 35, 559-586.

[21] Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. Humanities and Social Sciences Communications, 11, 723.

[22] Ma, R., Qian, M., Tang, S., et al. (2025). Assessment of L2 oral proficiency using speech large language models. In Proceedings of Interspeech 2025.

[23] Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. Journal of Second Language Writing, 68, 101203.

[24] Chan, S. T. S., Lo, N. P. K., & Wong, A. M. H. (2024). Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes. Contemporary Educational Technology, 16(4), ep541.

[25] Chapelle, C. A., Cotos, E., & Lee, J. Y. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language Testing, 32(3), 385-405.