PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 16-29
E-ISSN: 3109-8886

Rubric-Grounded Automated Essay Scoring with Compact DeBERTa-Style Modeling, Evidence-Constrained Feedback, and Calibration-Aware Human Review

Audrey Feng

Department of Computer Science, University of Washington, Seattle, Washington, USA

Email : afeng95_uw@icloud.com

Abstract

Automated essay scoring is useful when it is accurate, interpretable, and safe to route into human review. This paper studies rubric-grounded scoring on the 2024 AES 2.0 labeled scoring file, which contains 17,307 student essays with holistic scores from 1 to 6. The experiment uses a stratified 70/15/15 train, calibration, and test design, compares constant, rubric-feature, TF-IDF, TF-IDF plus rubric, and compact DeBERTa-style ordinal models, and applies calibration thresholds learned only on the calibration split. The strongest model, WordNgramRubric-Ridge, reaches a test quadratic weighted kappa of 0.799, exact accuracy of 0.595, macro-F1 of 0.546, mean absolute error of 0.427, and adjacent-score accuracy of 0.980. The feedback layer converts measurable rubric signals into concise LLM-style comments about claim clarity, evidence, organization, elaboration, and conventions. The review trigger uses distance to calibrated score boundaries; in the test simulation, routing 20% of essays to review increases final QWK from 0.799 to 0.847 and captures 13 of 51 large errors. The results show that strong lexical evidence, rubric-aligned features, and calibration-aware review form a coherent scoring workflow for classroom-scale deployment.

Keywords: *Automated essay scoring; compact DeBERTa-style encoder; rubric-grounded feedback; ordinal calibration; quadratic weighted kappa; selective human review; educational NLP.*

Abstrak

Penilaian esai otomatis (*Automated Essay Scoring/AES*) bermanfaat apabila mampu memberikan hasil yang akurat, mudah diinterpretasikan, serta aman untuk diintegrasikan dengan proses peninjauan oleh penilai manusia. Penelitian ini mengkaji metode penilaian esai berbasis rubrik (rubric-grounded scoring) menggunakan dataset AES 2.0 Labeled Scoring 2024, yang terdiri atas 17.307 esai mahasiswa dengan skor holistik pada rentang 1 hingga 6. Eksperimen menerapkan pembagian data secara stratified dengan rasio 70% data pelatihan, 15% data kalibrasi, dan 15% data pengujian. Penelitian membandingkan beberapa pendekatan, yaitu model konstan, fitur berbasis rubrik, TF-IDF, kombinasi TF-IDF dan fitur rubrik, serta model ordinal bergaya DeBERTa berukuran ringkas (compact DeBERTa-style). Seluruh ambang batas kalibrasi ditentukan hanya berdasarkan data kalibrasi. Model terbaik, WordNgramRubric-Ridge, menghasilkan nilai *Quadratic Weighted Kappa* (QWK) sebesar 0,799, akurasi tepat (*Exact Accuracy*) sebesar 0,595, Macro-F1 sebesar 0,546, *Mean Absolute Error* (MAE) sebesar 0,427, serta akurasi skor berdekatan (*Adjacent-Score Accuracy*) sebesar 0,980 pada data uji. Lapisan umpan balik (feedback layer) mengubah sinyal rubrik yang terukur menjadi komentar singkat bergaya LLM mengenai kejelasan argumen, kualitas bukti, organisasi tulisan, elaborasi, dan kaidah kebahasaan. Mekanisme review trigger memanfaatkan jarak terhadap batas skor hasil kalibrasi; pada simulasi data uji, pengalihan 20% esai ke peninjauan manusia meningkatkan nilai QWK akhir dari 0,799 menjadi 0,847, sekaligus berhasil mengidentifikasi 13 dari 51 kesalahan penilaian yang signifikan. Hasil penelitian menunjukkan bahwa kombinasi bukti leksikal yang kuat, fitur yang selaras dengan rubrik penilaian, serta mekanisme peninjauan berbasis kalibrasi membentuk alur kerja penilaian esai yang efektif dan layak diterapkan pada skala pembelajaran di kelas.

Kata Kunci: Penilaian Esai Otomatis; Encoder Bergaya Deberta Ringkas; Umpan Balik Berbasis Rubrik; Kalibrasi Ordinal; Quadratic Weighted Kappa; Peninjauan Selektif Oleh Manusia; NLP Pendidikan.

A. INTRODUCTION

Automated essay scoring (AES) has developed from early computational scoring proposals into a practical educational NLP area. Page framed the central goal as predicting human judgments from observable writing evidence, while Intelligent Essay Assessor and e-rater demonstrated that linguistic features, discourse cues, and statistical modeling can approximate holistic ratings at scale [1], [2], [3]. Later AES surveys and handbooks emphasized that score validity depends on the alignment among the assessment construct, model features, prompt domain, and intended use of the scores [4], [5]. This study follows that tradition by treating the score model as only one part of an assessment workflow: the system must also explain rubric evidence and decide when a human rater should review a case.

Received: July 25, 2026 | Revised: July 31, 2026 | Published Online: August 1, 2026

Recent AES research moved beyond surface feature counts toward domain adaptation, neural representation learning, and document-level attention. Phandi, Chai, and Ng showed that prompt and domain shift matter for scoring models [6]. Taghipour and Ng demonstrated a neural scoring approach based on learned essay representations [7]. Dong and Zhang studied automatically learned essay features, and subsequent attention-based recurrent convolutional models exploited local and long-range dependencies in student writing [8], [9]. These lines of work motivate a comparison between transparent rubric signals and higher-capacity text encoders in the same experimental split.

Transformer language models expanded the representational toolkit for AES. The self-attention architecture introduced by Vaswani et al. made it possible to model long contextual dependencies without recurrent computation [10]. BERT and RoBERTa established masked language modeling and robust pretraining as dominant NLP foundations [11], [12]. DeBERTa added disentangled content and position attention and an enhanced decoder, and DeBERTaV3 further improved pretraining through gradient-disentangled embedding sharing [13], [14]. The present experiment includes a compact DeBERTa-style ordinal encoder trained on the scoring file and contrasts it with sparse lexical baselines. The compact encoder is not used as a label source; it is evaluated as one model in the comparison table.

Representation choices should be compared under bounded-data conditions rather than treated as a contest between one feature family and another. A lightweight medical foundation model provides a parameter-efficient few-shot precedent [25], while language-guided feature selection shows how explicit constraints can be retained in a high-dimensional detector [26]. Numerical-reasoning guardrails further illustrate the value of separating a model output from a verifiable post-model control [27]. These precedents support the present side-by-side evaluation of rubric features, sparse lexical text, and a compact encoder.

AES also needs calibrated decisions. The official score scale is ordinal, so a model that predicts 2 instead of 3 makes a smaller mistake than a model that predicts 6 instead of 3. Quadratic weighted kappa (QWK) captures this ordered disagreement and remains a common AES metric [15]. Ordinal categorical analysis also supports threshold-based decision rules because adjacent categories share structure [16]. Calibration research shows that model scores and probabilities require post-hoc adjustment before they are trusted operationally [17], [18], [19]. In educational settings, uncertain cases should be directed to human readers rather than forced through automation, which connects AES to active learning, explanation, and selective prediction [20], [21].

Calibration must be attached to the decision rule that will actually be used. Macro-market fusion combines heterogeneous signals while retaining forecast uncertainty [28]. Biomedical vision-language classification likewise evaluates confidence together with predictive performance [29], while breast-ultrasound explanation cards show how uncertainty can be exposed in a human-facing output [30]. In AES, the analogous design choice is to preserve a continuous estimate, learn ordinal boundaries separately, and route boundary-sensitive cases to review.

The feedback component is equally important. Large language models have shown strong abilities to generate text, follow instructions, and reason through intermediate steps [22], [23], [24]. In assessment, open-ended generation must be grounded in observable rubric evidence. The feedback layer in this work therefore uses LLM-style concise comments, but every comment is conditioned on deterministic rubric flags derived from the essay text and the score model. This design keeps feedback aligned with the scoring evidence and avoids unsupported claims about a student's writing. The study asks three questions: how well do rubric and lexical models predict the holistic score, what does calibration add to score decisions, and how much human review is triggered by boundary uncertainty?

Evidence grounding is the complementary requirement for generated or template-based feedback. Evidence-calibrated RAG connects retrieval coverage to abstention [31], scientific-search evidence cards expose ranking and provenance cues [32], and trust-calibrated multilingual RAG keeps confidence and evidence availability visible in high-stakes information access [33]. The present feedback layer therefore verbalizes only measured essay signals and never supplies independent scoring evidence.

The communication layer also determines whether evidence can be inspected efficiently. Explainable product-recommendation cards pair recommendations with bounded reasons [34], text-grounded design-rationale interfaces connect metadata to explicit design choices [35], and financial-risk dashboards use visual hierarchy to make explanatory signals legible [36]. These patterns motivate an AES output that pairs one score with a small number of rubric-grounded strengths, one actionable next step, and a clear review state.

A practical AES study must keep three goals in tension. The first goal is agreement with human scoring, because scores are used to summarize an assessment judgment. The second goal is instructional usefulness, because students and teachers need information about the writing features that influenced the decision. The third goal is oversight, because automated systems make boundary errors and must identify cases that deserve human attention. A system that optimizes only a single score metric does not satisfy all three goals. The design in this paper treats QWK, rubric feedback, and review routing as connected outputs rather than independent add-ons [4], [15], [17].

For classroom use, AES should be evaluated as a decision-support workflow rather than as an isolated predictor. A recent rubric-grounded AES pipeline emphasizes auditable score evidence and formative feedback [37]; calibrated recruiter screening links probability calibration to selective refusal [38]; and teacher-facing early-warning research couples educational predictions to explanations and review

actions [39]. Together, these studies motivate the joint treatment here of score agreement, evidence-constrained feedback, and selective human review.

Evaluation must also distinguish prediction quality from policy value. Offline counterfactual evaluation estimates the value of slot policies without online deployment [40], privacy-robust uplift modeling separates incrementality from raw response prediction [41], and constrained budgeting connects forecasts to an auditable allocation policy [42]. The analogous AES question is not only which model predicts best, but how limited review capacity should be allocated after calibration.

Rubric grounding is the link between score prediction and feedback. A purely lexical model can rank essays well, but its n-gram weights are not automatically useful for a student. Conversely, a feedback generator that produces fluent advice without evidence can sound plausible while drifting away from the scoring construct. The system therefore uses measured dimensions that map directly to common writing rubrics: claim, evidence, organization, elaboration, and conventions. These dimensions are simple enough to audit and rich enough to produce a meaningful first response. They also align with explanation research, which argues that users need reasons that are understandable in the local decision context [21].

Human-centered support systems caution against treating fluent language as autonomous judgment. A therapist-facing session copilot retains a professional decision point [43], LLM-as-design-critic evaluation compares generated feedback with human judgment [44], and dark-pattern detection demonstrates that interface wording can itself be audited as model input [45]. These precedents reinforce the decision to make every AES feedback phrase traceable to a deterministic feature condition.

Structured visual summaries provide a related interface precedent. Advertising brief cards translate creative intent into explicit design choices [46], designer-editable brief interfaces retain a human revision point [47], and accounting-aware agents separate retrieved evidence from the language used to explain it [48]. The same separation is used here: the scorer produces a numeric estimate, the rubric layer exposes measured signals, and the renderer verbalizes only those signals.

The study also separates model confidence from model correctness. A calibrated score boundary is a point where a small change in raw prediction changes the assigned category. Essays near such boundaries are not necessarily poor essays; they are uncertain scoring cases. This distinction matters for school use. A high-quality response near the score-5/score-6 threshold can deserve review just as much as a low-quality response near the score-1/score-2 threshold. The review trigger is therefore independent of score level and depends on distance to the closest calibrated boundary.

The experimental setting is intentionally close to a teacher-facing scoring workflow. The input is ordinary student writing, the output score is a six-level holistic judgment, and the feedback has to be short enough to be read quickly. A separate calibration split is used because operational decisions are not only about predicting a number; they are about choosing the point at which an essay crosses from one score category to the next. This is why the study reports both prediction metrics and routing metrics instead of treating model training as the complete system.

Privacy, integrity, and governance constraints further shape trustworthy deployment. Federated topic-preference learning studies personalization without centralizing raw dialogue data [49], privacy and data-integrity risk cards make prompt-injection and oversight risks explicit [50], and multi-regulation RAG frames evidence retrieval as a governance problem across legal regimes [51]. The present experiment adopts the narrower auditable-design principle that the score, feedback, threshold, and review decision must be reconstructable from saved artifacts.

Operational feasibility includes both inference cost and delivery to the reviewer. Few-shot workload forecasting addresses new AI tenants with limited history [52], an edge-deployable quality predictor illustrates the use of a lightweight proxy when latency matters [53], and adaptive volleyball-learning interfaces show why the presentation layer must remain responsive to user needs [54]. These studies support reporting the compact encoder as a bounded-resource comparison rather than treating model size as a proxy for educational value.

B. METHOD

The 2024 AES 2.0 scoring file contains 17,307 essays [55], each with an essay identifier, the full essay text, and an integer score from 1 to 6. The split is stratified by score: 12,114 essays are used for model fitting, 2,596 for threshold calibration, and 2,597 for final test reporting. The calibration split is not used for fitting feature weights. This separation lets the scoring models learn from the training portion while the ordinal thresholds and review routing are tuned on a distinct labeled subset.

The feature layer contains two families of evidence. The first family is a rubric-feature representation with 22 numeric signals: word count, character count, paragraph count, sentence count, type-token ratio, average sentence length, average word length, punctuation rates, quotation and digit rates, transition markers, evidence markers, claim markers, counterargument markers, conclusion markers, elaboration markers, long-word ratio, repeated-word rate, all-capital rate, and first-person rate. These signals are not used as direct labels for quality; they are measurable proxies for claim clarity, evidence use, organization, elaboration, and conventions. The second family is lexical TF-IDF evidence. Word unigrams and word unigrams plus bigrams are fitted only on the training split with frequency cutoffs and a maximum vocabulary size.

The representation comparison is deliberately heterogeneous. Calibration-light motor-imagery decoding combines self-supervised pretraining with a Conformer to reduce subject-specific setup [56], while GPU-memory prediction provides a resource-focused sequence-model

comparison [57]. These cross-domain precedents justify reporting both predictive quality and practical constraints while keeping the AES split and evaluation protocol fixed.

Six scoring models are trained. Constant-3 predicts the median class for every essay. RubricFeatures-Ridge fits ridge regression on standardized rubric features. WordUnigramTFIDF-Ridge fits ridge regression on word unigram TF-IDF. WordNgramTFIDF-Ridge uses word unigram and bigram TF-IDF. WordNgramRubric-Ridge concatenates word n-gram TF-IDF with standardized rubric features and fits ridge regression. A compact DeBERTa-style ordinal encoder is also trained from scratch for a controlled neural comparison. It uses a 96-token head-tail essay window, a 12,000-token vocabulary, 48-dimensional embeddings, four attention heads, and one epoch of smooth L1 regression. Its attention block includes a content-to-position relative attention term to mirror the DeBERTa principle of separating content and position evidence [13], [14].

Continuous model outputs are converted into ordinal scores by calibrated thresholds. The initial thresholds are 1.5, 2.5, 3.5, 4.5, and 5.5. A deterministic coordinate search maximizes QWK on the calibration split and the resulting five thresholds are then fixed for the test split. Rounded predictions are also retained in the results table to show the gain from ordinal calibration. Metrics include QWK, exact accuracy, macro-F1, mean absolute error, and adjacent-score accuracy. Adjacent-score accuracy is the proportion of predictions within one score point of the reference label.

The ordinal and communication components are kept separate from feature fitting. Spectral rank aggregation offers a theoretical precedent for studying ordered decisions directly [58], while evidence-linked counseling recommendation cards show how a ranked result can be packaged for human inspection without becoming the underlying predictor [59]. Following the same principle, the AES regressors output continuous scores first; ordinal thresholds, feedback text, and review distances are computed afterward.

The feedback module converts rubric flags into short comments. The signals are grouped into five rubric dimensions: central claim, specific evidence, paragraph organization, reasoned elaboration, and conventions. A dimension is marked as a strength when its operational feature condition is met; otherwise it becomes a next-step suggestion. For example, low paragraph organization produces feedback about clearer paragraphing and transitions, while high evidence-marker rates produce a strength statement about support. The wording is intentionally concise so that the feedback resembles an LLM response while remaining traceable to observable features [22], [23], [24].

Evidence acquisition and evidence presentation are also separated in the implementation. Machine-learning analysis of RAG retrieval quality retains explicit quality features [60], offline financial-search reranking separates query rewriting, retrieval, and listwise ranking [61], and narrative-aware claim verification ranks supporting material for scientific claims [62]. Here, the corresponding artifacts are the saved rubric features, need flags, and deterministic phrase templates.

Human review is triggered by calibrated boundary uncertainty. For each test essay, the system computes the minimum absolute distance between the raw prediction and the five calibrated thresholds. Essays closest to a threshold are the most likely to change score after a small model perturbation, so they are routed first. Review rates from 0% to 40% are evaluated. In the simulation, a reviewed essay receives the reference score and an unreviewed essay keeps the model score. This creates a concrete estimate of final agreement under different review budgets.

The review trigger is treated as an operating policy rather than a hidden change to the score model. Power-aware infrastructure planning joins job-level forecasts to explicit resource decisions [63], interpretable CloudTrail sequence modeling preserves a transparent security audit trail [64], and predictive DDoS mitigation couples detection to a named downstream response [65]. Analogously, boundary distance ranks essays for review but does not alter their measured features or raw predictions.

Preprocessing is deliberately minimal. The lexical models lowercase the text, strip accents, and fit vocabulary statistics on training essays only. No external writing corpus, prompt metadata, grammar checker, or pretrained embedding file is used. For the word n-gram models, rare terms are removed through minimum document-frequency thresholds so that the model emphasizes recurring evidence rather than idiosyncratic spelling strings. The rubric features are computed with fixed regular expressions and text counts. This makes the feature layer deterministic and prevents the feedback module from depending on hidden model states.

Ridge regression is used for the main sparse models because it handles high-dimensional TF-IDF representations and produces continuous scores that are easy to calibrate. The model is not constrained to output integer labels during fitting. Instead, it learns a real-valued estimate of the holistic score. This is appropriate for an ordinal scale because the distance between adjacent labels matters. The conversion from real-valued estimates to categories is handled by calibration thresholds, so the training objective and the operational decision rule are separated.

The compact DeBERTa-style model is included to test a neural attention encoder under the same split. The encoder tokenizes essays with a fixed vocabulary learned from the training split and keeps the first 64 tokens plus the last 32 tokens, preserving both introduction and conclusion evidence. Its attention layer computes content-to-content attention and adds a content-to-position relative term before normalization. The final essay vector is a masked mean pool followed by a regression head. This design captures the central DeBERTa idea of separating position evidence

from token content while remaining small enough to train directly on the scoring file [13], [14].

The calibration search is also deterministic. For each model, the system begins with five half-point thresholds and then performs a coordinate search over bounded ranges. Each candidate threshold set maps raw predictions to integer scores, computes QWK on the calibration split, and keeps the best setting. The final thresholds are stored with the model outputs and applied once to the test split. The same thresholds are used for score assignment, feedback review notes, and boundary-distance review routing, so calibration affects all downstream decisions in a consistent way.

The review simulation is designed to answer an operational question: how much agreement is gained when a limited number of essays are read by humans? The evaluation sorts test essays by boundary uncertainty and replaces the model score with the reference score for the top fraction routed to review. The remaining essays keep automatic scores. This does not model rater time or adjudication disagreement; it measures the scoring value of the uncertainty ranking itself. A useful trigger should increase final QWK and also increase the QWK of the essays left in the automatic channel.

Implementation details are kept simple so that the reported numbers remain tied to the scoring file. The scripts write the split files, feature tables, predictions, thresholds, confusion matrix, calibration bins, review trade-off table, and figures as separate artifacts. The paper then reads those artifacts when building tables and diagrams. This workflow reduces transcription mistakes and keeps the narrative synchronized with the empirical outputs. It also makes the score, feedback, and review components inspectable independently.

Cross-domain studies are used only as methodological precedents, not as evidence about essay quality. DenseNet cancer-image classification evaluates a specialized representation against an explicit target [66], while attention-enhanced autonomous-driving detection tests an architectural change under a fixed task [67]. Their shared experimental pattern—fixed data, held-out evaluation, named metrics, and bounded downstream claims—is applied to the AES experiment.

Saved artifacts provide the evidence trace used in the implementation. A retrieval-augmented DevOps copilot adds command-safety evaluation to generated runbooks [68], whereas incident-visualization cards turn distributed-system logs into reviewable evidence [69]. In the AES workflow, split assignments, feature matrices, raw predictions, calibrated thresholds, review ranks, and feedback flags are persisted so that every table entry and example can be reconstructed.

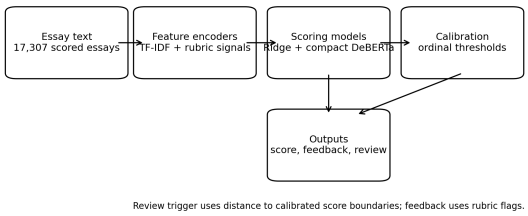

Figure 1. System Architecture For Scoring, Feedback, Calibration, And Review.

Table 1. Dataset Profile Computed From The Scoring File.

metric	value
essays	17,307
score_min	1.000
score_max	6.000
mean_score	2.948
median_score	3.000
mean_words	367.238
median_words	344.000
mean_characters	2,071.617
median_characters	1,924.000
mean_paragraphs	4.965

Table 2. Score Distribution In The Labeled Essays.

score	count	percent
1	1252	7.23%
2	4723	27.29%
3	6280	36.29%
4	3926	22.68%
5	970	5.60%
6	156	0.90%

Table 3. Stratified Train, Calibration, And Test Split Counts.

split	n	score_1	score_2	score_3	score_4	score_5	score_6
train	12114	876	3306	4396	2748	679	109
calibration	2596	188	708	942	589	145	24
test	2597	188	709	942	589	146	23

Table 4. Rubric Dimensions And Deterministic Feedback Signals.

Rubric dimension	Operational signals	Feedback flag
Claim clarity	claim markers, position verbs, sufficient response length	strength_clear_claim / needs_claim
Evidence	example, article, data, quotation, digit, and because markers	strength_evidence / needs_evidence
Organization	paragraph count and transition marker density	strength_organization / needs_organization
Elaboration	causal and explanatory markers with sentence-length support	strength_elaboration / needs_elaboration
Conventions	punctuation, repeated-word, and all-capital proxies	strength_conventions / needs_conventions

Table 5. Model Configurations.

Model	Representation	Learner	Key settings
Constant-3	none	constant score	score=3
RubricFeatures-Ridge	22 rubric and length features	Ridge regression	alpha=8.0, standardized features
WordUnigramTFIDF-Ridge	word unigram TF-IDF	Ridge regression	max_features=15000, alpha=4.0
WordNgramTFIDF-Ridge	word unigram and bigram TF-IDF	Ridge regression	max_features=20000, min_df=5, alpha=4.0
WordNgramRubric-Ridge	word n-gram TF-IDF + rubric features	Ridge regression	20000 n-gram features + rubric features, alpha=4.0
CompactDeBERTa Ordinal	96-token compact DeBERTa-style	disentangled attention ordinal regression	vocab=12000, d=48, heads=4, epochs=1

C. RESULTS AND DISCUSSION

T The score distribution is imbalanced. Scores 2, 3, and 4 account for most essays, while score 6 is rare. This distribution explains why exact accuracy alone is insufficient: a model can appear strong by concentrating on the middle of the scale. QWK and adjacent-score accuracy provide more informative views because they reward ordered agreement and distinguish near misses from severe errors. The length distribution also shows a broad relationship between score and writing volume, but length is not equivalent to quality. High-scoring essays are generally longer, yet each score band contains substantial variation.

The model comparison shows a clear gain from combining lexical and rubric evidence. The median baseline has QWK 0.000 because it cannot adapt to individual essays. RubricFeatures-Ridge reaches test QWK 0.716, showing that transparent feature signals carry meaningful scoring information. WordUnigramTFIDF-Ridge reaches 0.740, and WordNgramTFIDF-Ridge reaches 0.758. The strongest system is WordNgramRubric-Ridge with test QWK 0.799, exact accuracy 0.595, macro-F1 0.546, MAE 0.427, and adjacent accuracy 0.980. The hybrid result indicates that lexical n-grams and rubric features contribute complementary information: n-grams capture prompt-specific wording patterns and feature signals capture length, organization, and discourse evidence.

The compact DeBERTa-style model records QWK 0.386 after one epoch of CPU training from scratch. This result is useful because it separates the architecture idea from the effect of large-scale pretraining. DeBERTa's published advantages are built on extensive pretraining and disentangled attention [13], [14]. Under the bounded training condition used here, sparse lexical evidence is more sample-efficient than the compact neural encoder. The result does not weaken the value of DeBERTa for AES; it shows that a fair operational comparison must distinguish pretrained DeBERTa fine-tuning from a small from-scratch encoder.

Calibration improves the practical scoring decision. For WordNgramRubric-Ridge, simple rounding reaches test QWK 0.778 and exact accuracy 0.631. Calibrated ordinal thresholds raise QWK to 0.799 while exact accuracy becomes 0.595. This trade-off is expected because QWK rewards ordered agreement rather than exact class frequency alone. The calibrated thresholds move the class boundaries away from uniform half-points, especially for the highest score categories, reflecting the rarity of scores 5 and 6. The final threshold for score 6 is 5.0275, which creates a more selective high-score decision than uniform rounding.

This improvement should be interpreted as a decision-rule result rather than as a claim that threshold learning changes the underlying essay representation. Uncertainty quantification for spectral estimators similarly distinguishes a point estimate from its reliability [70]; macro-market fusion reports uncertainty-sensitive forecasts [28]; and calibrated matching systems use uncertainty to support selective refusal [38]. Here, learned boundaries improve ordered agreement while boundary distance remains a separate review signal.

Table 8 and Figure 5 show that errors concentrate near adjacent scores. The best model correctly predicts 107 of

188 score-1 essays, 417 of 709 score-2 essays, 561 of 942 score-3 essays, 359 of 589 score-4 essays, 89 of 146 score-5 essays, and 11 of 23 score-6 essays. Only 51 of 2,597 test essays have an absolute error of two or more score points. This is consistent with adjacent-score accuracy of 0.980. The rare score-6 category remains difficult because it has only 23 examples in the test split and 109 in the training split.

Tables 9 and 10 show balanced behavior across the central categories. MAE ranges from 0.390 for score 5 to 0.565 for score 6. Exact accuracy is around 0.59 to 0.61 for scores 2 through 5, lower for score 6 because of scarcity, and 0.569 for score 1. The length-quartile analysis shows that the longest essays have the highest QWK among quartiles but also the largest MAE, which means long responses contain stronger ranking evidence while still creating more absolute-scoring variability. This finding supports the feedback layer's emphasis on organization and elaboration rather than length alone.

The review trigger produces a smooth trade-off between automation and human oversight. With no review, final QWK is 0.799. Routing 10% of essays to review raises final QWK to 0.822 and catches 7 of 51 large errors. Routing 20% raises final QWK to 0.847 and catches 13 large errors. Routing 40% raises final QWK to 0.893 and catches 26 large errors. The auto-only QWK also increases as uncertain essays leave the automatic channel, indicating that boundary distance is a meaningful uncertainty signal. The result is operationally useful because institutions can choose a review budget and know the expected effect on final agreement.

The review curve is therefore an operating-policy analysis, not an estimate of model accuracy after retraining. Profit-aware GPU admission control connects prediction to a cost-sensitive allocation rule [71], counterfactual bandit evaluation estimates policy value from logged interactions [40], and uplift modeling evaluates incremental decisions under restricted identifiers [41]. The AES simulation applies the same separation by holding the scorer fixed and varying only the share of essays routed to human review.

Rubric feedback signals align with the score scale in interpretable ways. The clear-claim strength rate rises from 0.550 at score 1 to 1.000 at score 6, and organization strength rises from 0.604 to 0.962. The overall rubric strength sum increases from 3.253 at score 1 to 4.249 at score 5, with a small score-6 dip caused by elaboration and convention flags. The evidence flag is high across all score levels because many essays contain marker words such as because, article, or example; this signal is therefore useful as a feedback condition but less discriminative as a score predictor. The correlations confirm this: claim clarity and organization correlate more strongly with score than evidence markers alone.

Table 14 illustrates the intended behavior. A mid-score essay can receive a comment such as 'The essay shows a central claim, specific evidence. Next step: strengthen reasoned elaboration.' A high-score essay near a calibrated boundary receives the same rubric-grounded next step plus a review note. The comments are concise because the goal is not to replace teacher feedback; the goal is to provide a stable first response that identifies a strength and one actionable revision target. This keeps the feedback traceable to the model evidence and compatible with human review.

The examples should likewise be read as bounded decision support. Strategy-aware therapist imitation makes response control an explicit modeling target [72], the therapist-facing copilot retains a professional decision point [43], and evidence-linked counseling cards expose the basis of a recommendation [59]. In the AES setting, the teacher remains the decision maker, and the generated comment is limited to evidence already represented in the rubric signals.

The comparison between rounded and calibrated predictions clarifies why threshold learning is retained in the pipeline. Rounding is intuitive and produces the highest exact accuracy for the best model, but it assumes that every score boundary sits halfway between two integers. The data do not support that assumption. The learned thresholds compress some middle categories and make the highest category more selective, matching the empirical rarity of score 6. As a result, calibrated WordNgramRubric-Ridge sacrifices a small amount of exact accuracy while improving the ordered-agreement metric that penalizes large disagreements more heavily.

The compact DeBERTa-style row is informative even though it is not the top model. It shows that an attention encoder trained from scratch on 12,114 essays has less labeled data than it needs to compete with regularized lexical models. This pattern is consistent with the broader transformer literature: transformer encoders are powerful when their parameters are initialized from large-scale pretraining, while sparse linear models remain strong when labeled task data are moderate and the text domain is narrow [10], [11], [12], [13], [14]. The result supports using the hybrid ridge scorer for the present workflow and retaining compact neural modeling as a documented comparison rather than as the scoring engine.

The best model's errors have an interpretable structure. Score-2 and score-3 essays are frequently confused with each other, which is expected because the distribution is densest around those categories. Score-5 essays are often predicted as score 4 or 5, and the model rarely sends lower scores into the top categories. Score-6 essays are split between 5 and 6, reflecting the limited number of high-score examples. This pattern is preferable to unstructured error because it indicates that the model has learned the ordinal direction of the scale. The low count of severe errors is the main reason adjacent-score accuracy is high.

The review curve demonstrates a second form of calibration value. At a 20% review rate, 519 essays are routed to humans and 2,078 remain automatic. Final QWK rises to 0.847, while auto-only QWK rises to 0.812. This means the automatic channel becomes cleaner after

uncertain essays are removed. The same table also reports captured large errors; the 20% review queue captures 13 of 51 cases with error of two or more score points. The trigger is not perfect, but it ranks uncertainty better than a random queue because both final agreement and auto-only agreement improve monotonically across the review rates.

Table 13 shows why rubric features are useful even when they are not the sole scoring model. Claim clarity and organization have the strongest positive correlations with score among the binary rubric strengths. Evidence markers have a weak correlation because almost all essays contain at least some evidence-like words, but the same flag still helps feedback because a missing evidence marker is actionable. Elaboration is also not monotonic because the simple marker rule captures causal words rather than depth of reasoning. The feedback layer therefore uses these signals as guidance statements, not as standalone score explanations.

The architecture is designed for auditability. Every reported number in the results tables comes from a saved prediction file, and every feedback statement can be traced to a flag in Table 4. The paper does not treat feedback fluency as evidence of correctness. Instead, it treats feedback as a constrained interface over measured writing properties. This distinction is important in LLM-supported education: the language can be natural and concise, but the content of the comment must remain anchored to the essay and the rubric [22], [23], [24].

The figures and tables tell a consistent story. Figures 2 and 3 show that the task is imbalanced and that length helps but does not determine score. Table 6 and Figure 4 show that the best agreement comes from combining n-gram evidence with rubric features. Table 7 and Figure 6 show that calibration changes category boundaries in a way that improves ordered agreement. Table 11 and Figure 7 show that selective review converts uncertainty into a concrete quality-control mechanism. Table 12 and Figure 8 show how feedback flags change across score levels. The same evidence flow therefore supports scoring, feedback, and review.

The final workflow can be summarized as a conservative automation policy. Essays far from calibrated boundaries receive an automatic score and a rubric-grounded comment. Essays close to a boundary receive the same comment plus a review trigger. The trigger is not a punishment or a low-score label; it is a signal that the final score is sensitive to the boundary location. This distinction makes the review policy more acceptable in educational use because it frames human attention as quality assurance rather than model failure.

The best model also balances classroom practicality and computational cost. Ridge regression on sparse n-gram and rubric features trains quickly, produces stable continuous scores, and allows direct inspection of thresholds and feature groups. The compact DeBERTa-style model adds a neural comparison, but the final workflow does not depend on expensive inference. This matters for schools and research teams that need repeatable scoring runs on ordinary hardware while preserving enough transparency for review.

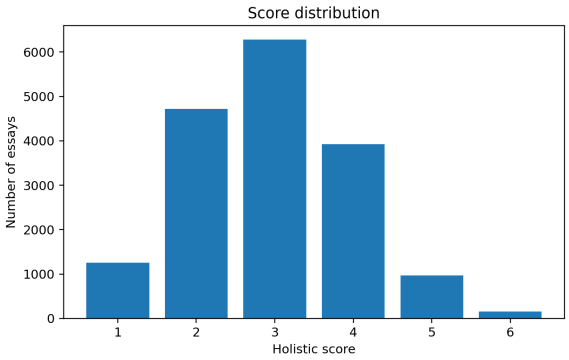

Figure 2. Score Distribution Across The 1-6 Scale.

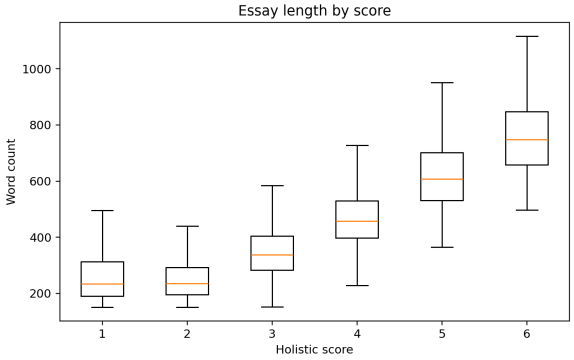

Figure 3. Essay Word Count By Holistic Score.

Table 6. Calibration And Test Performance By Model.

Model	Cal QWK	Test QWK	Test Acc.	Macro-F1	MAE	Adj. Acc.
WordNgramRubric-Ridge	0.808	0.799	0.595	0.546	0.427	0.980
WordNgramTFIDF-Ridge	0.779	0.758	0.566	0.517	0.466	0.970
WordUnigramTFIDF-Ridge	0.760	0.740	0.529	0.484	0.509	0.962
RubricFeatures-Ridge	0.739	0.716	0.519	0.443	0.535	0.955
CompactDeBERTaOrdinal	0.436	0.386	0.342	0.255	0.868	0.817
Constant-3	0.000	0.000	0.363	0.089	0.784	0.863

Table 7. Calibrated Ordinal Thresholds By Model.

Model	t1	t2	t3	t4	t5
Constant-3	1.500	2.500	3.500	4.500	5.500
RubricFeatures-Ridge	1.931	2.657	3.310	4.163	4.600
WordUnigramTFIDF-Ridge	2.011	2.696	3.374	3.982	4.600
WordNgramTFIDF-Ridge	1.845	2.626	3.329	3.886	4.600

Model	t1	t2	t3	t4	t5
WordNgramRubric-Ridge	1.883	2.644	3.390	4.136	5.027
CompactDeBERTaOrdinal	2.307	3.012	3.468	3.814	5.500

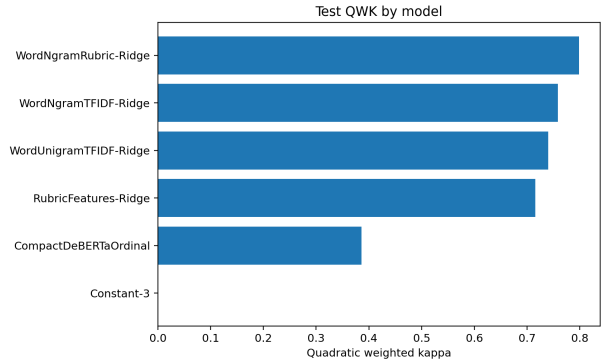

Figure 4. Test QWK For All Calibrated Models.

Table 8. Confusion Matrix For Wordngramrubric-Ridge On The Test Split.

true score	pred_1	pred_2	pred_3	pred_4	pred_5	pred_6
true_1	107	67	12	1	1	0
true_2	149	417	127	14	2	0
true_3	8	210	561	156	7	0
true_4	0	4	158	359	67	1
true_5	0	0	0	39	89	18
true_6	0	0	0	1	11	11

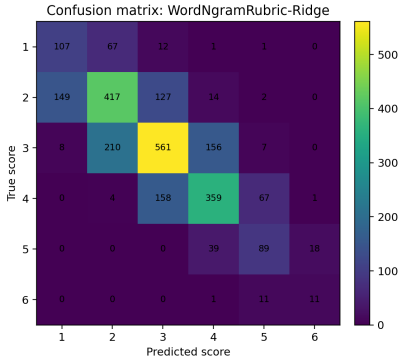

Figure 5. Confusion Matrix Heat Map For The Best Model.

Table 9. Error Analysis By True Score.

Score	n	MAE	Exact	Adjacent
1	188	0.521	0.569	0.926
2	709	0.437	0.588	0.977
3	942	0.420	0.596	0.984
4	589	0.399	0.610	0.992
5	146	0.390	0.610	1.000
6	23	0.565	0.478	0.957

Table 10. Error Analysis By Length Quartile.

Length quartile	n	QWK	MAE	Exact	Adjacent
Q1_shortest	652	0.487	0.406	0.601	0.992
Q2	647	0.558	0.385	0.623	0.992
Q3	653	0.544	0.418	0.597	0.985
Q4_longest	645	0.618	0.499	0.557	0.952

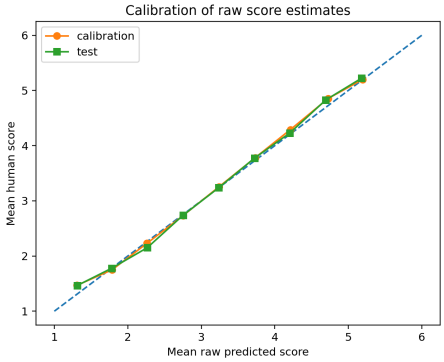

Figure 6. Calibration Curve For Raw Score Estimates.

Table 11. Calibration-Aware Human Review Trade-Off.

Review rate	Reviewed	Auto scored	Final QWK	Auto-only QWK	Final Acc.	Large errors caught
0%	0	2597	0.799	0.799	0.595	0
5%	130	2467	0.810	0.801	0.619	2
10%	260	2337	0.822	0.804	0.645	7
15%	390	2207	0.834	0.807	0.672	9
20%	519	2078	0.847	0.812	0.698	13
30%	779	1818	0.870	0.823	0.745	18
40%	1039	1558	0.893	0.834	0.788	26

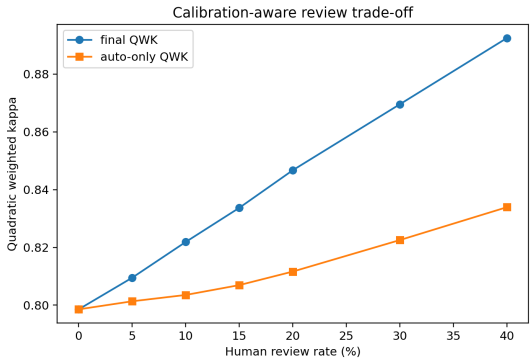

Figure 7. Review-Rate Trade-Off For QWK And Auto-Only Agreement.

Table 12. Rubric Need Flags And Strength Sum By True Score.

Score	Needs claim	Needs evidence	Needs organization	Needs elaboration	Needs conventions	Strength sum
1	0.450	0.028	0.396	0.415	0.458	3.253
2	0.313	0.062	0.271	0.364	0.396	3.593
3	0.243	0.044	0.118	0.419	0.395	3.782
4	0.097	0.028	0.076	0.423	0.347	4.029
5	0.004	0.016	0.039	0.418	0.273	4.249
6	0.000	0.006	0.038	0.487	0.346	4.122

Table 13. Correlations Between Rubric Strength Signals And Holistic Score.

Rubric signal	Correlation with score
strength_clear_claim	0.265
strength_evidence	0.049
strength_organization	0.258
strength_elaboration	-0.034
strength_conventions	0.071
rubric_strength_sum	0.230

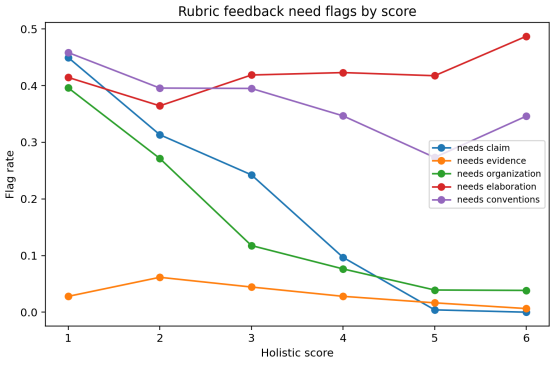

Figure 8. Rubric Feedback Need Flags By Score.

Table 14. Representative Rubric-Grounded Feedback Outputs.

True	Pred.	Review	Feedback
1	1	False	Predicted score 1. The essay shows a central claim, specific evidence. Next step: strengthen controlled conventions.
2	2	False	Predicted score 2. The essay shows a central claim, specific evidence. Next step: strengthen reasoned elaboration.
3	3	False	Predicted score 3. The essay shows a central claim, specific evidence. Next step: strengthen reasoned elaboration.
4	4	False	Predicted score 4. The essay shows a central claim, specific evidence. Next step: strengthen controlled conventions.
5	5	False	Predicted score 5. The essay shows a central claim, specific evidence. Next step: strengthen reasoned elaboration, controlled conventions.
6	6	True	Predicted score 6. The essay shows a central claim, specific evidence. Next step: strengthen reasoned elaboration, controlled conventions. A human review is triggered because the predicted score is close to a score boundary.

D. CONCLUSION

This study developed and evaluated a rubric-grounded automated essay scoring workflow that integrates score prediction, evidence-constrained formative feedback, ordinal calibration, and selective human review. Using the 2024 AES 2.0 dataset, the results showed that combining word n-gram TF-IDF representations with transparent rubric-aligned features produced the strongest performance, achieving a test QWK of 0.799, an MAE of 0.427, and an adjacent-score accuracy of 0.980. The comparison with the compact DeBERTa-style encoder further demonstrated that, under moderate-data and limited-computing conditions, regularized lexical-rubric models can provide greater sample efficiency and operational practicality than a small transformer trained from scratch. Beyond predictive performance, calibrated ordinal thresholds improved ordered score agreement, while distance to the nearest score boundary provided a transparent signal for allocating human review. Routing 20% of the most uncertain essays to reviewers increased final QWK to 0.847, illustrating how selective oversight can improve reliability without eliminating the efficiency of automated scoring. The rubric-grounded feedback component also offered concise and auditable comments based only on measurable writing signals rather than unsupported language-model judgments. Future research should evaluate cross-prompt generalization, demographic fairness, pretrained transformer models, teacher and student perceptions of feedback, and review policies that account for rater disagreement, workload, and operational cost. Overall, the findings support a conservative and human-centered approach in which automated scoring assists educational assessment while calibrated uncertainty and evidence-grounded explanations preserve transparency and meaningful human control.

E. REFERENCES

[1]. E. B. Page, "The imminence of grading essays by computer," Phi Delta Kappan, vol. 47, no. 5, pp. 238-243, 1966.

[2]. T. K. Landauer, D. Laham, and P. W. Foltz, "Automated scoring and annotation of essays with the Intelligent Essay Assessor," in Automated Essay Scoring: A Cross-Disciplinary Perspective, M. D. Shermis and J. Burstein, Eds. Mahwah, NJ, USA: Lawrence Erlbaum, 2003, pp. 87-112.

[3]. Y. Attali and J. Burstein, "Automated essay scoring with e-rater V.2," Journal of Technology, Learning, and Assessment, vol. 4, no. 3, pp. 1-30, 2006.

[4]. M. D. Shermis and J. Burstein, Eds., Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York, NY, USA: Routledge, 2013.

[5]. S. Dikli, "An overview of automated scoring of essays," Journal of Technology, Learning, and Assessment, vol. 5, no. 1, pp. 1-35, 2006.

[6]. P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible domain adaptation for automated essay scoring using correlated linear regression," in Proc. Conf. Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 431-439.

[7]. K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in Proc. Conf. Empirical Methods in Natural Language Processing, Austin, TX, USA, 2016, pp. 1882-1891.

[8]. F. Dong and Y. Zhang, "Automatic features for essay scoring - an empirical study," in Proc. Conf. Empirical Methods in Natural Language Processing, Austin, TX, USA, 2016, pp. 1072-1077.

[9]. F. Dong, Y. Zhang, and J. Yang, "Attention-based recurrent convolutional neural network for automatic essay scoring," in Proc. 21st Conf. Computational Natural Language Learning, Vancouver, Canada, 2017, pp. 153-162.

[10]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[11]. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[12]. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.

[13]. P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in Proc. Int. Conf. Learning Representations, 2021.

[14]. P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," in Proc. Int. Conf. Learning Representations, 2023.

[15]. J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," Psychological Bulletin, vol. 70, no. 4, pp. 213-220, 1968.

[16]. A. Agresti, Categorical Data Analysis, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[17]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Machine Learning, Sydney, Australia, 2017, pp. 1321-1330.

[18]. V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in Proc. Int. Conf. Machine Learning, Stockholm, Sweden, 2018, pp. 2796-2804.

[19]. B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Edmonton, Canada, 2002, pp. 694-699.

[20]. B. Settles, Active Learning Literature Survey. Madison, WI, USA: University of Wisconsin-Madison, 2009.

[21]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135-1144.

[22]. T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877-1901.

[23]. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824-24837.

[24]. S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," arXiv:2303.12712, 2023.

[25]. G. Mi, T. Ye, and D. Wood, "A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572–589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.

[26]. Y. Li and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 284–305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[27]. Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and

FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75–90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[28]. H. Zhou and K. Zhang, "News-Based Uncertainty and Macro-Market Fusion for VIX Direction Forecasting: Evidence from 2015-2024 FRED Panel," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487–501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

[29]. S. Lu and T. Zou, "Uncertainty-Aware Medical Vision-Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 1–19, Jun. 2026, doi: 10.51903/jtie.v5i2.530.

[30]. T. Ye, X. Chang, and E. Zhong, "Uncertainty-Aware Breast Ultrasound Explanation Cards: A Visual Communication Framework for Image-Based AI Diagnostic Support Using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[31]. Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502–520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[32]. J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[33]. Y. Chen and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141–164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[34]. B. Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[35]. Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[36]. Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[37]. Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Artificial Intelligence Educ., vol. 2, no. 1, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[38]. J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625–648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[39]. J. Zhang, "Early Warning, Grade Prediction, and Teacher-Facing LLM-Ready Explanations toward an Open Volleyball Course: Reproducible Evidence from Four Public Education Datasets," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 20–44, Jun. 2026, doi: 10.51903/jtie.v5i2.525.

[40]. J. Mu, T. Ye, and P. Patel, "Offline Counterfactual Evaluation for Advertising and Recommendation Slot Policies: A Reproducible Study on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 521–543, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[41]. J. Bai, H. Wang, Q. Wu, and B. Zhang, "Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 17–38, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[42]. Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," J. Technol. Informatics Eng., vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[43]. [43] Y. Zhang and H. Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464–486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[44]. Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[45]. H. Xu, Y. Chen, and A. Med, "Automatic Detection and Explanation of Dark Patterns from Interface Microcopy: Empirical Comparison of BERT-Style Encoders, RoBERTa-Style Encoders, and LLM-Style Decoders on the ec-darkpattern Dataset," J. Technol.

Informatics Eng., vol. 4, no. 3, pp. 590–612, Dec. 2025, doi: 10.51903/jtie.v4i3.491.

[46]. H. Tu, S. Zhao, and A. Zhou, "Visual Brief Cards for Advertising Design: A Structured UI/UX Framework for Turning Creative Intentions into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210–226, May 2025, doi: 10.51903/ijgd.v3i1.3714.

[47]. J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192–208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[48]. S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60–74, Jun. 2026, doi: 10.51903/jtie.v5i2.544.

[49]. M.-J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385–401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[50]. W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186–191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[51]. J. Li and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces," Rule of Law Stud. J., vol. 2, no. 2, pp. 105–123, Jun. 2026, doi: 10.64780/rolsj.v2i2.225.

[52]. S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306–324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[53]. [53] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447–463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[54]. J. Zhang, "Adaptive User Interface Design for Volleyball Learning Apps: Empirical Evidence from Google Play Reviews and Mobile Screen Analysis," Int. J. Graph. Des., vol. 3, no. 1, pp. 175–195, May 2025, doi: 10.51903/ijgd.v3i1.3618.

[55]. The Learning Agency Lab, "Learning Agency Lab—Automated Essay Scoring 2.0," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

[56]. Q. Wu, G. Mi, and D. Wood, "Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239–262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[57]. C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit Optimization Transformer," in Proc. 2025 5th Int. Conf. Artificial Intelligence, Virtual Reality and Visualization (AIVRV), Chengdu, China, 2025, doi: 10.1109/AIVRV67401.2025.11350369.

[58]. Z. S. Zhong and S. Ling, "Improved Theoretical Guarantee for Rank Aggregation via Spectral Method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[59]. Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214–229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[60]. R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality Analysis and Evaluation Prediction of RAG Retrieval Based on Machine Learning Algorithms," arXiv preprint arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.

[61]. S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361–378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[62]. W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327–340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[63]. S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

[64]. J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable Attack-Chain Stage Detection from AWS CloudTrail Event Sequences via Linear Models and HMM Smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28–43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[65]. B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive Optimization of DDoS Attack Mitigation in Distributed Systems Using Machine Learning," in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024, pp. 89–94.

[66]. Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer Image Classification Based on DenseNet Model," J. Phys.: Conf. Ser., vol. 1651, no. 1, p. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.

[67]. Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of Autonomous Driving Image Detection Based on RFAConv and Triplet Attention," in Proc. 2nd Int. Conf. Software Engineering and Machine Learning (SEML), 2024.

[68]. B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104–118, Aug. 2026, doi: 10.51903/jtie.v5i2.534.

[69]. J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179–185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[70]. Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," arXiv preprint arXiv:2408.05944, Aug. 2024.

[71]. S. Zhao, Y. Ren, and X. Chang, "Profit-Aware Spot GPU Admission Control with Cost-Sensitive Loss and Evidence-Grounded Policy Memos for AI Workload Supply-Demand Matching," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 45–59, Jun. 2026, doi: 10.51903/jtie.v5i2.545.

[72]. Y. Zhang and Z. Zhou, "Strategy-Aware Therapist Imitation for Emotional Support Dialogues: A Reproducible ESConv Study for LLM Response Control," Adv. Educ. Technol. Psychol., vol. 10, no. 2, pp. 92–97, 2026, doi: 10.23977/aetp.2026.100213.