scientific reports

Check for updates

OPEN

A hybrid intelligent assessment model for English translation education with improved BERT and SVM

Chuan Lin

Assessing student translations in real classrooms still leans heavily on human judgment, which can vary across raters, is time-consuming, and scales poorly. With the rapid advancement of natural language processing (NLP), automatic assessment models have shown potential to enhance objectivity and consistency in translation evaluation. This paper proposes BERT-SVM EduScore, a hybrid assessment model designed for English translation education. The model couples a BERT-base encoder adapted with domain-/task-specific training and a contrastive objective with compact linguistic and alignment features, and feeds the resulting representations into margin-based Support Vector Machines (SVM) heads with monotonic calibration to produce holistic and rubric-aligned sub-scores. We conduct experiments on the English–Chinese portion of the MLQE-PE machine-translation quality estimation dataset, which we use as a proxy for sentence-level scoring in classroom-like settings. BERT-SVM EduScore is compared against string-based metrics (BLEU, METEOR, TER, chrF++), embedding-based semantic metrics (YiSi-1, MoverScore, BERTScore), and learned evaluators (BLEURT, COMET, PRISM, BARTScore, TransQuest) using Quadratic Weighted Kappa (QWK), mean absolute error (MAE), Pearson's r, and runtime. On MLQE-PE it achieves QWK 0.76, MAE 0.12, and r 0.84, improving QWK over COMET by + 0.08 and over TransQuest by + 0.05 while running at about 22.5 ms per sentence (≈ 44 sentences per second) on commodity GPU hardware. Ablation studies show that domain-adaptive pretraining yields the largest gains, contrastive learning provides additional improvements, and metric distillation contributes smaller but consistent benefits. These quantitative results suggest that the proposed model can serve as a technically feasible component for translation education, offering calibrated scores and lightweight diagnostic signals under realistic latency constraints; validation on real course assignments and with teachers and students in the loop is left for future work.

Keywords English translation education, BERT, Support vector machine (SVM), Hybrid intelligent model

Assessing student translations is inherently challenging because rater judgments are subjective, rubrics vary across instructors and institutions, and the limited time available for feedback in classroom settings restricts the level of detail that can be provided[1,2]. These factors reduce inter-rater reliability and can make scores unstable across different evaluators. To address these challenges, standardized analytic schemes such as the Multidimensional Quality Metrics (MQM) framework have been introduced. MQM formalizes error types, severities, and scoring procedures, thereby promoting consistency in evaluation and providing a more transparent basis for instructional use[3]. Alongside human-driven schemes, the field has relied on automatic metrics such as BLEU, METEOR, and chrF to enable scalable assessment of translations[4–6]. These metrics are computationally efficient and reproducible, but they emphasize surface-level n-gram overlap and often penalize legitimate paraphrases, lexical variations, and alternative syntactic structures that are common in learner language. As a result, their diagnostic value in educational settings is limited.

Correlations between these string-based metrics and human judgments are also sensitive to reference diversity and task setup, which further motivates the development of semantically informed evaluation methods that go beyond string matching[7]. Representation-based metrics that exploit contextual embeddings, including

Academic Journal Editorial Office, Zhengzhou University of Technology, Zhengzhou 450044, Henan, China. email: Chuanlin19@outlook.com

BERTScore, COMET, and BLEURT, demonstrate stronger alignment with human assessment at both segment and system levels compared with traditional baselines[8,9,11]. The importance of semantic evaluation has been reinforced through large-scale shared tasks: the WMT Metrics task quantifies the degree of agreement between metrics and human ratings, while the WMT Quality Estimation task evaluates reference-free prediction at word, sentence, and document levels[12,13]. These benchmarks provide rigorous testbeds and highlight both the promise and the remaining challenges of automated metrics.

In the context of translation education, instructors need not only reliably scoring but also diagnostic and actionable feedback that can be used to guide students. They also require stability when working with small cohorts, yet these needs are not fully met by purely end-to-end neural metrics under data constraints[1,3]. Pre-trained Transformers such as BERT offer transferable representations that can be adapted to educational domains, capturing semantic adequacy and error cues more effectively for assessment[10]. Support Vector Machines (SVM) contribute complementary strengths, offering large-margin robustness and strong generalization in high-dimensional feature spaces, which is particularly advantageous when labeled data are scarce—a typical condition in classroom scenarios[14]. Hybrid pipelines that combine Transformer-based embeddings with shallow classifiers such as SVM have been shown to achieve competitive accuracy and stability, suggesting a principled and practical path toward reliable scoring in translation education[15].

In practical classroom settings, however, instructors often face a substantial marking load and strict time constraints, which limit the amount and timeliness of detailed feedback they can provide on individual sentences. Scores may drift across raters, cohorts, and semesters, especially when rubrics are complex and when classes are taught by multiple instructors or teaching assistants. Existing automatic metrics are useful for benchmarking systems, but they typically provide only a single holistic score without clear diagnostic cues and are not always efficient enough for real-time use in teaching platforms. The concrete educational problem we address is therefore how to provide stable, calibrated, and diagnostically informative sentence-level scores under the constraints of limited data and modest computational resources, so that instructors can use them as support for formative feedback and moderation rather than as a replacement for human judgment.

Accordingly, we propose a hybrid intelligent assessment model that improves BERT via education-oriented domain adaptation and employs an SVM head to deliver robust, fine-grained quality judgments on limited data. We evaluate against strong baselines (BLEU, METEOR, chrF) and learned metrics (BERTScore, COMET, BLEURT) and validate with course-level human ratings to quantify agreement, reliability, and diagnostic utility. The main contributions are outlined below.

(1) This paper introduces BERT-SVM EduScore framework, an improved BERT scorer stacked with SVM/SVR, to output calibrated holistic and sub-dimension scores.

(2) We unify reference-free QE and reference-based evaluation via compact linguistic/alignment features and provide classroom-ready explanations (terminology/number alignment; SHAP/LIME).

(3) We show on cross-topic/cross-student splits that our model outperforms strong baselines on QWK/MAE/Pearson while maintaining low-latency inference for classroom use.

This paper is organized as follows: "Related work" section reviews the literature and formalizes the problem. "Methodology" section presents the model. "Experiments" section describes the datasets, setup, and results, with ablation and efficiency analyses. "Conclusion" section addresses limitations and ethics and concludes.

Related work

Reference-based MT metrics and limitations

SacreBLEU highlighted how pre-processing variance distorts BLEU comparability and urged standardized scoring pipelines to ensure like-for-like reporting across tokenization, casing, and segmentation choices[16,17]. Building on the limits of word n-grams, chrF++ added word-level n-grams atop character n-grams to better capture morphology and orthographic variation, though it remains fundamentally overlap-centric and thus sensitive to reference sparsity[18]. Learned or reference-free trends were anticipated by Prism, which leverages a multilingual paraphraser to compare meaning rather than strings and thereby reduces penalties for valid paraphrases in low-reference or cross-lingual setups[19]. Representation-driven metrics such as MoverScore and BARTScore push further toward semantics by using contextual embeddings and sequence-to-sequence likelihoods, respectively, improving segment-level alignment with human judgments beyond surface overlap[20,21].

YiSi[22] unifies semantic evaluation and estimation across resource conditions and languages, offering a consistent scaffold when references are scarce or heterogeneous. Large-scale WMT Metrics shared tasks from 2019 to 2022 consistently showed neural/semantic metrics correlating more strongly with human judgments than n-gram overlap metrics at both system and segment levels, reinforcing the field's migration beyond pure string matching[23–25]. At the same time, these neural metrics can be compute-intensive and sensitive to domain shift and reference diversity, motivating designs that combine semantically informed encoders with lightweight decision heads and calibrated outputs for reliability in real-world settings such as translation education, where actionable feedback, stable scoring with small cohorts, and classroom-grade latency are paramount[17–25].

Semantic and learned metrics

Beyond overlap, semantic metrics operationalize meaning by comparing representations induced by pretrained encoders or by scoring sequences with generative language models. BERTScore aligns hypothesis–reference pairs in embedding space (often with IDF weighting and layer selection) to emphasize semantic adequacy, while BLEURT fine-tunes a pretrained encoder as a learned regressor against human ratings, improving sensitivity to fluency and adequacy cues beyond surface matches. Complementary advances include Prism's multilingual zero-shot paraphrasing to canonicalize content before comparison, MoverScore's Earth-Mover (Wasserstein) distance over contextualized token embeddings to capture flexible alignments, BARTScore's use of sequence log-likelihood under a pretrained seq2seq model to reflect plausibility and adequacy from the model's generative

prior, and YiSi's unified design spanning evaluation and estimation across languages and resource regimes, which helps when references are scarce or heterogeneous[19–22].

Evidence from WMT21/22 shows that neural/semantic metrics are more robust than BLEU-style baselines across domains (news, TED/social/e-commerce/chat) and at both system- and segment-level granularity which often under MQM-based human evaluation, supporting a shift toward representation-driven judgments[24,25]. At the same time, these metrics can be compute-intensive and occasionally brittle under domain shift or reference sparsity, motivating hybrid designs that retain semantic sensitivity while improving stability, calibration, and efficiency for education-oriented deployment.

Quality estimation (QE)

Quality estimation (QE) predicts translation quality without references, matching classroom settings where curated gold references are rare and heterogeneous. QuEst + + formalized multi-level QE as a feature-engineering and learning pipeline spanning sentence and word granularity, establishing reproducible baselines and error-aware features[26]. OpenKiwi operationalized this line in PyTorch, packaging winning WMT15–18 systems with consistent data loaders, training recipes, and evaluation at word/sentence levels, thereby enabling fair comparisons and rapid iteration[27].

With stronger encoders, TransQuest introduced cross-lingual Transformer architectures that learn shared semantic spaces across languages, delivering competitive sentence-level QE and favorable transfer to low-resource pairs where labeled judgments are scarce[28]. To support modern evaluation, MLQE-PE unified multilingual QE with sentence-level direct assessment, word-level tags, and post-editing signals, standardizing splits and metrics for recent shared tasks and facilitating joint modeling across granularities[29]. Results from WMT21 QE further consolidated encoder-based and multilingual setups, highlighting that reference-free, representation-driven models can provide stable ranking and actionable signals (e.g., error spans/tags) suitable for educational triage, formative feedback, and moderation workflows where scalability and consistency are essential[24].

Domain/task adaptation and contrastive learning

Domain-adaptive pretraining (DAPT/TAPT) improves transfer to in-domain text under both high- and low-resource settings, which is critical for education corpora that diverge from newswire[30]. Parameter-efficient adapters enable task-specific tuning with few extra parameters, preserving a frozen backbone for stability and efficiency[31]. SBERT and SimCSE provide high-quality sentence embeddings via siamese/triplet and contrastive learning, strengthening segment-level semantic similarity essential for evaluation and QE[32,33]. In our setting, DAPT/TAPT aligns the encoder to learner-language distributions, adapters permit rapid per-course retargeting without catastrophic forgetting, and a contrastive objective (inherited from SBERT/SimCSE) sharpens sentence-level discrimination and yielding data-efficient representations that pair well with shallow SVR heads and calibration for robust classroom scoring[30–33].

Hybrid framework and explainability for education

In data-limited settings, pairing frozen Transformer embeddings with lightweight heads such as linear regression models or SVM and SVR classifiers has been shown to provide accuracy that is competitive with more complex neural architectures. This strategy also preserves robustness by limiting overfitting and controlling inference latency, which is especially important in classroom environments where computational resources are modest and annotated data are scarce. Recent BERT–SVM variants demonstrate strong performance on semantic matching and related classification tasks that share structural similarities with translation assessment, and they can be deployed efficiently on CPUs without requiring specialized hardware. The combination of pretrained embeddings and margin-based predictors also makes it easier to apply calibration methods, which leads to stable and interpretable score distributions that align with rubrics used in language education[26–28,34].

For actionable classroom feedback, model-agnostic explanation methods help make system outputs transparent and pedagogically useful. Such explanations clarify how individual tokens, features, or linguistic cues contribute to the final score, which enables instructors to validate system behavior and students to understand their mistakes. Anchors can generate high-precision local rules that highlight decisive lexical or structural features and are straightforward for instructors to audit. Integrated Gradients provides axiomatic attributions that distribute importance scores across tokens, which is suitable for modern text models. These techniques are particularly valuable for surfacing terminology mismatches, number and unit inconsistencies, or punctuation errors that matter in translation grading. By linking explanations directly to rubric categories, they also support human-in-the-loop moderation, allowing instructors to intervene when the automatic system shows low confidence or inconsistent reasoning. In this way, the combination of lightweight prediction heads and model-agnostic explanations offers a practical balance of efficiency, reliability, and interpretability for translation assessment in educational settings[34,35].

Methodology

The proposed methodology is a hybrid translation-assessment framework that domain/task-adapts BERT with a contrastive objective, fuses pooled embeddings with compact linguistic diagnostics, and outputs calibrated holistic and sub-dimension scores via margin-based SVR (optional banding and reference-aware distillation), achieving high agreement at classroom-level latency in both reference-based and reference-free modes.

The framework is designed with English translation education in mind, with the long-term goal of supporting automated formative feedback on assignments, rubric-aligned summative scoring and moderation, placement or diagnostic testing, and class-scale analytics. In this paper, however, we focus on the technical feasibility of the scoring component and evaluate it offline on a public quality-estimation corpus as a proxy for classroom grading. Latency and throughput experiments indicate that the model could be deployed on commodity CPU/

GPU hardware, but actual integration with learning management systems and systematic user studies with teachers and students are left for future work.

Framework overview

This section introduces BERT-SVM EduScore, a hybrid framework for classroom translation assessment that supports both reference-free quality estimation (QE) and reference-based scoring. Given a source x, a student translation h, and an optional reference r, an improved BERT encoder produces a sentence-level representation that is fused with compact linguistic/alignment features; lightweight SVR heads yield holistic and rubric-aligned sub-scores, while an optional SVM head outputs categorical bands or error types. Training combines domain-adaptive pretraining and a contrastive objective for representation quality, metric distillation from strong learned metrics when references exist, and monotonic calibration to align predictions with instructor scales[3,31,35]. The end-to-end pipeline is summarized in Fig. 1.

Problem formulation and notation

An improved BERT encoder produces sentence-level representations, which are concatenated with compact linguistic/alignment features to form a unified feature vector for lightweight SVR/SVM heads[10,14,22]. It can be written as Eq. (1).

$$\mathrm{z} = \mathrm{Pool}\left(\mathrm{BERT}\left([x; h; r]\right)\right), u = [\mathrm{z}; \mathrm{f}], \hat{y} = g_{SVR}^{(K)}(\mathrm{u}) \tag{1}$$

where, we assess a student translation h for a source sentence x with an optional reference r_t , predicting a holistic score $\widehat{y} \in \mathbb{R}$ and sub-dimension scores $\widehat{y} \in \mathbb{R}^K$ aligned with instructional rubrics (e.g., adequacy, fluency, terminology) under both reference-free QE and reference-based settings[3,12,13].

The diagram of improved BERT encoder is shown in Fig. 2.

Encoder adaptation: DAPT and contrastive learning

To mitigate domain mismatch (learner language, terminology, common errors), we perform domain-adaptive pretraining (DAPT/TAPT) on classroom corpora[30], optionally with parameter-efficient adapters for stability and efficiency[31]. To strengthen sentence-level semantics and error sensitivity, we add a SimCSE/SBERT-style contrastive loss over augmented positives (paraphrases/back-translations) and hard negatives (minimal edits)[32,33]. The detail of $\mathcal{L}_{ctr}$ is shown in Eq. (2).

$$\mathcal{L}_{ctr} = -\sum_{i} \log \frac{\exp\left(\cos\left(z_i, z_i^{+}\right)/\tau\right)}{\sum_{j \in \mathcal{N}(i)} \exp\left(\cos\left(z_i, z_j\right)/\tau\right)} \tag{2}$$

where, τ is a temperature, z_i^{+} denotes the positive of anchor i, and $\mathcal{N}(i)$ contains positives and negatives[32,33]. The algorithm of encoder adaptation is shown in Table 1.

Our improved BERT encoder preserves the standard BERT-base Transformer architecture. We do not modify the pretraining objectives as in RoBERTa, introduce cross-layer parameter sharing as in ALBERT, or compress the model as in DistilBERT. Instead, we retain the publicly available BERT-base checkpoints and adapt them to the translation-education setting via domain- and task-adaptive pretraining, optional parameter-efficient adapter

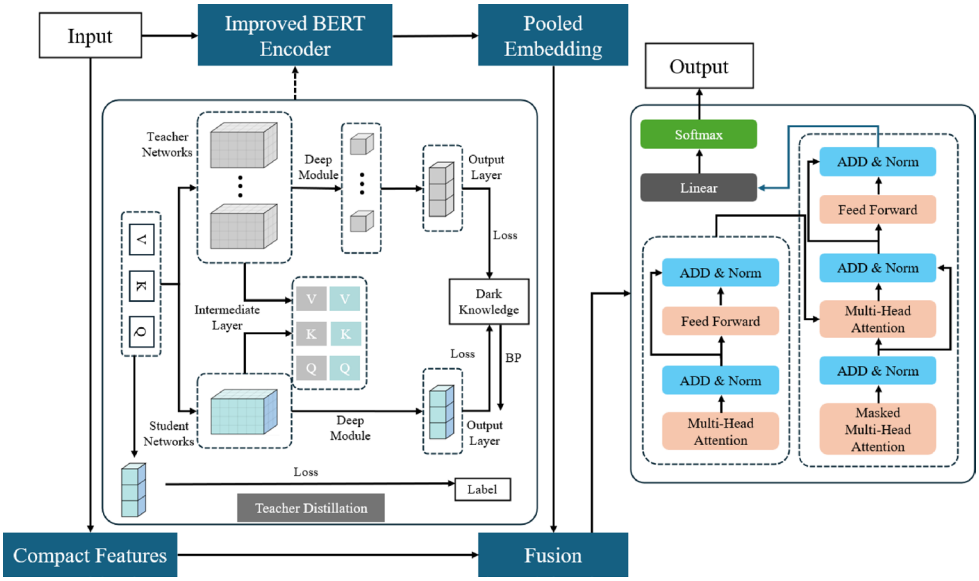

Fig. 1. BERT–SVM EduScore framework overview.

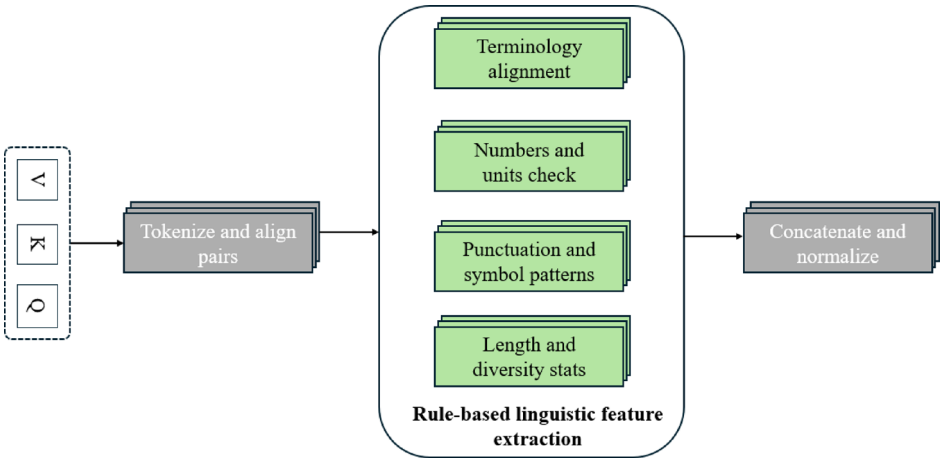

Fig. 2. Improved BERT encoder.

Data: U, (x, h), r, τ, B, K, λ_{mlm} , λ_{ctr}
Result: E_θ
initialize encoder: $E_\theta \leftarrow$ pretrained BERT; insert adapters; optionally freeze lower layers
for each epoch do
  sample mini-batch $\mathcal{B}_{text}$
  compute L_{mlm}
  sample mini-batch $\mathcal{B} = \{(x_i, h_i, [r_i])\}\{i = 1..B\}$
  for $i = 1, B$ do
    $P_i \leftarrow (x_i, h_i)$
    $N_i \leftarrow$ minimal-edit hard negatives of (x_i, h_i)
    $z_i \leftarrow E_\theta$ (concat(x_i, h_i, r_i))
    $z_i^+ \leftarrow E_\theta$ (sample(P_i))
    $Z_{ineg} \leftarrow \{ E_\theta (n) \mid n \in N_i$
  end for
  $L_{\text{ctr}} \leftarrow (1/B) \sum_i$ NT-Xent$(z_i, z_i^+, Z_{ineg}$; τ)
  Ltotal $\leftarrow \lambda_{mlm} \cdot L_{mlm} + \lambda_{ctr} \cdot L_{\text{ctr}}$
  update θ with Adam
end for
return E_θ

Table 1. Algorithm of encoder adaptation.

layers, and the SimCSE/SBERT-style contrastive objective described above. This design keeps the architecture simple and well understood while tailoring the learned representations to learner-like input and the scoring task.

Algorithm 1 summarizes these steps. Given a mini-batch of sentence pairs, we first apply domain-adaptive pretraining updates on masked language modeling examples drawn from the in-domain corpus. We then construct contrastive pairs by sampling augmented positives and hard negatives, compute the contrastive loss over normalized sentence embeddings, and combine it with the task loss when human labels are available. Gradients are backpropagated through the encoder (and adapter layers, when used), while the SVR/SVM heads are trained on pooled features in a subsequent stage. This separation keeps the encoder adaptation procedure modular and allows us to reuse the same encoder across different prediction heads.

Compact linguistic and alignment features

We compute a compact yet informative set of diagnostic features designed to complement neural representations while keeping inference overhead minimal. These features include terminology hit-rate, number and unit consistency, punctuation discipline, length ratio, lexical diversity, language-ID confidence, and simple alignments to the reference when it is available[3,12,22]. Each of these captures a distinct aspect of translation quality that is frequently emphasized in instructional rubrics. For example, terminology and number checks highlight factual accuracy, punctuation discipline reflects basic fluency, and lexical diversity provides a signal of stylistic adequacy. Language-ID confidence helps detect code-switching or off-task outputs, while length ratio serves as

Data: x, h, r, G, U, $\{min_j, max_j\}$, small $\varepsilon > 0$
Result: $f \in R^7$, [0,1]
$tokens_x \leftarrow tokenize(x)$; $tokens_h \leftarrow tokenize(h)$
$f1_{term} \leftarrow hit_{rate}(terms(tokens_x, G), h)$
$f2_{num} \leftarrow consistency(numbers_{units}(x, U), numbers_{units}(h, U))$
$f3_{punc} \leftarrow punctuation_{discipline}(h)$
$f4_{len} \leftarrow \lvert tokens_h \rvert / max(\lvert tokens_x \rvert, 1)$
$f5_{ttr} \leftarrow \lvert unique(tokens_h) \rvert / max(\lvert tokens_h \rvert, 1)$
$f6_{lid} \leftarrow lid_{confidence}(h, target_{lang})$
$f7_{ref} \leftarrow charF(h, r)$ *if r exists else* 0
for $j = 1..7$ *do* f[j] $\leftarrow clip((f_{raw}[j] - min_j) / max(max_j - min_j, \varepsilon), 0, 1)$
return f

Table 2. Algorithm of feature extraction.

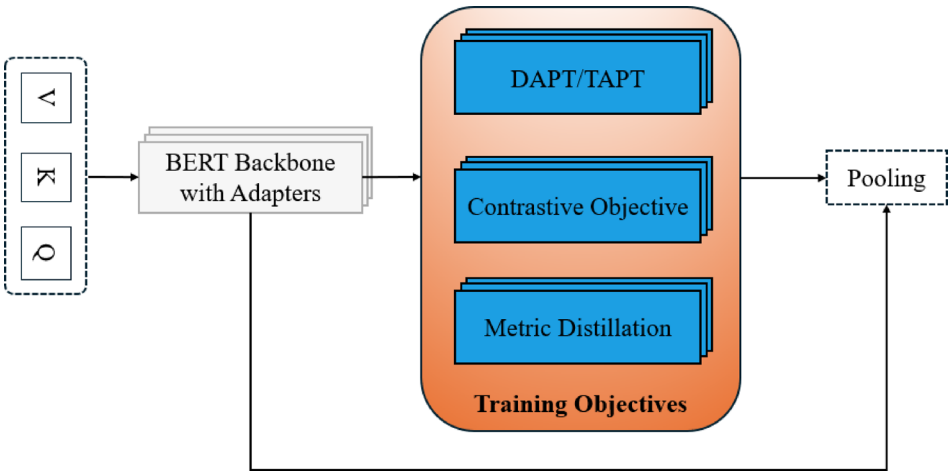

Fig. 3. Feature extraction.

a coarse proxy for coverage of source content. By combining these lightweight signals with deep representations, the framework strengthens robustness in cases where the encoder alone may be misled by paraphrases or rare constructions.

All feature values are min–max normalized on the training data to provide consistent scaling and then clipped to the interval [0,1] during inference to suppress outliers. This normalization ensures that no single feature dominates the decision process and that the features can be fused seamlessly with pooled embeddings. The procedure is efficient and requires only simple token-level operations, which guarantees that it can be executed in parallel with encoder inference without adding noticeable latency. The step-by-step algorithm for feature extraction is presented in Table 2, and the corresponding flowchart illustrating the extraction pipeline is provided in Fig. 3, which together clarify how linguistic cues are computed, normalized, and integrated into the hybrid scoring model.

A detailed feature-importance analysis for these compact diagnostics is beyond the scope of the present paper. While our qualitative inspection suggests that terminology, number consistency, punctuation discipline, and length ratio align well with common rubric categories, a more systematic study using feature ablations or attribution methods (e.g., SHAP) on larger and more diverse datasets is left for future work.

Prediction heads and calibration

Continuous holistic and sub-dimension scores are produced by SVR on u; when rubrics require categorical bands or error types, an SVM head is added and jointly trained[14]. We then learn isotonic (default) or Platt calibration Φ on a development set to align predictions with instructor scales across cohorts[12,13]. The SVR objective and calibration mapping are shown in Eq. (3) and Eq. (4) as

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i (\xi_i + \xi_i^*) \text{ s.t. } \begin{cases} y_i - \left(w^{\mathrm{T}} u_i + b\right) \le \varepsilon + \xi_i, \\ \left(w^{\mathrm{T}} u_i + b\right) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \end{cases} \quad (3)$$

$$\tilde{y} = \Phi(\hat{y}), \Phi = arg \min_{isotonic\Phi} \sum_i (\Phi(\hat{y}_i) - y_i)^2 \quad (4)$$

where, $u_i = \in \mathbb{R}^{d+m}$ denotes the unified feature vector of sample i; $y_i = \in \mathbb{R}$ denotes its human score (holistic or a specific sub-dimension); w and b denote the SVR parameters; $\varepsilon > 0$ denotes the ε-insensitive margin; $C > 0$ denotes the regularization constant balancing flatness and violations; $\xi_i, \xi_i^* \geq 0$ denote slack variables for positive/negative residuals; $\mathcal{K}(\cdot,\cdot)$ denotes the kernel function (RBF by default, $\mathcal{K}\left(u, u'\right) = exp\left(-\gamma \| u - u' \|^2\right)$ with $\gamma > 0$. For multi-output scoring, Eq. (3) is optimized independently for each dimension $k \in \{1, \ldots, K\}$ with parameters $\left(w^{(k)}, b^{(k)}\right)$ and targets $y_i^{(k)}$.

Multi-objective training and metric distillation

To align predictions more closely with human judgments while still leveraging strong learned signals when available, we adopt a composite training objective that integrates multiple complementary components. The first part is a mean squared error (MSE) regression loss, which anchors the model outputs to continuous human ratings and encourages stable calibration. The second part is a pairwise ranking loss, which enforces relative ordering between better and worse translations within a batch, thereby improving robustness when absolute scores are noisy. In scenarios where reference translations are available, we further apply metric distillation by aligning the model outputs with scores produced by strong external teachers such as COMET and BLEURT[9,11,25,33]. This step allows the model to inherit semantic sensitivity from established metrics and improves generalization when human labels are sparse.

To avoid overfitting to teacher biases or unreliable signals, we introduce a confidence gate that suppresses the distillation term when the teacher predictions are detected as out-of-distribution relative to the training domain. This gating mechanism ensures that the student model benefits from teacher knowledge only when the additional signal is trustworthy, maintaining consistency with human labels as the primary target. The complete training objective thus combines regression, ranking, and selective distillation in a balanced form, ensuring that the model captures both absolute quality levels and relative preferences. The total objective is formally defined in Eq. (5), which integrates these terms with tunable weights to control their relative contributions during optimization.

$$\mathcal{L} = \lambda_m \frac{1}{N} \sum_i (\hat{y}_i - y_i)^2 + \lambda_r \sum_{(i,j)} \log\left(1 + \exp\left[-s_{ij}(\hat{y}_i - \hat{y}_j)\right]\right) + \lambda_k \frac{1}{N} \sum_i (\hat{y}_i - t_i)^2 + \lambda_c \mathcal{L}_{ctr} \quad (5)$$

here, $s_{ij} = \text{sign}(y_i - y_j)$, t_i is the teacher score, and λ are tuned on validation[9,11]. The algorithm of Supervised Training is shown in Table 3. The diagram of COMET and BLEURT is shown in Fig. 4.

Inference and explanations

At test time, we encode the input triple (x, h, r) compute the compact feature vector f, obtain predictions from the SVR or SVM heads, and then apply the calibration function ϕ to generate stable and rubric-aligned scores. For pedagogical feedback, we further compute token-level attributions with Integrated Gradients and extract Anchors rules, which allows evidence to be mapped explicitly to rubric dimensions such as adequacy, fluency, and terminology[34,35]. With the encoder frozen and tokenization cached, inference cost is dominated by a single forward pass through the Transformer and the evaluation of linear or RBF heads. Adapter-based tuning and parameter-efficient updates keep both memory consumption and latency modest, making the system deployable in real classroom settings[14,31]. Beyond numeric outputs, the explanations also support diagnostic guidance for students by highlighting specific lexical or structural issues that triggered deductions, and they allow instructors to audit the model's rationale and override decisions if necessary. The combination of calibrated scoring and

for minibatch $(x, h, [r], y)$ do
  $z \leftarrow BERT([x; h; r])$; $f \leftarrow$ features(x, h, r); $u \leftarrow [z; f]$
  $y_{hat} \leftarrow SVR(u)$; (opt) bands $\leftarrow$ SVM(u)
  $L_{mse} \leftarrow MSE(y_{hat}, y)$
  $L_{rank} \leftarrow pairwise_{rank}(y_{hat}, y)$
  if r:
    $t \leftarrow TeacherMetric(x, h, r)$
    $w \leftarrow KD_{gate}(t)$
    $L_{kd} \leftarrow w * MSE(y_{hat}, t)$
    $L \leftarrow \lambda_m \cdot L_{mse} + \lambda_r \cdot L_{rank} + \lambda_k \cdot L_{kd} + \lambda_c \cdot L_{ctr}$
  update encoder/heads
  end

Table 3. Algorithm of supervised training.

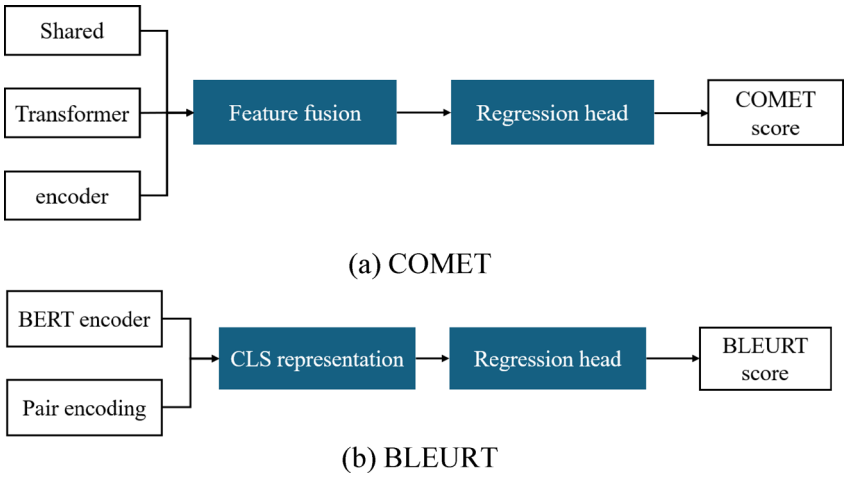

Fig. 4. Teacher metric as COMET or BLEURT.

interpretable feedback provides a dual benefit: it enables scalable automatic evaluation while still aligning with the instructional goals of fairness, transparency, and actionable error analysis. In practical use, this design allows the framework to operate not only as a grading assistant but also as a formative tool that encourages learners to reflect on their translation strategies and correct recurring mistakes.

Experiments

Datasets and setup

We adopt MLQE-PE (Multilingual Quality Estimation and Automatic Post-editing) as the single corpus for training and evaluation. Each instance contains a source sentence, a machine translation hypothesis, an optional human reference, and sentence- or word-level quality-estimation labels (e.g., direct assessment scores, HTER), which enables both reference-free QE and reference-based supervision within one protocol[36]. Although MLQE-PE consists of machine translations rather than student work, its sentence-level human scores are similar in form to instructor ratings of translation quality, so we use it as a proxy for classroom scoring in this study. We do not claim that MLQE-PE fully reflects the distribution of course assignments or institution-specific rubrics; extending the evaluation to real student translations with local scoring schemes is an important direction for future research.

Text is normalized (Unicode NFC), tokenized with WordPiece, and length-clipped with source-aware budgets; punctuation and numeric tokens are preserved to support diagnostic features. We employ stratified cross-student and cross-topic splits; development folds are disjoint from test folds and used for calibration and hyper-parameter selection via cross-validation with random search[37,38].

The encoder is BERT-base with parameter-efficient adapters. Prediction heads are SVR for holistic and sub-dimension scores and an optional SVM for categorical bands; hyper-parameters are selected by random search with early stopping on QWK[37,38].

The model is implemented in PyTorch 1.12 with CUDA 11.6 and HuggingFace Transformers; SVM/SVR components use scikit-learn. Experiments run on a single NVIDIA RTX 3090 (24 GB VRAM). We adopt BERT-base with lightweight adapters; inputs are tokenized by WordPiece and truncated/padded to 128 tokens per view ($QE : [CLS]\, x\, [SEP]\, h; [CLS]\, h\, [SEP]\, r$), with tri-view sequences capped at 256 tokens. Encoder adaptation (DAPT + contrastive) uses mixed precision (FP16), batch size 64, Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1e-8$), weight decay 0.01, base learning rates of $2 \times 10–5$ for the encoder and 1×10^{-3} for adapters/heads, layer-wise LR decay 0.95, 10% linear warm-up followed by cosine decay, and gradient clipping at 1.0. The contrastive objective employs temperature $\tau = 0.07$ with a 1:1 ratio of positives (paraphrase/back-translation) to hard negatives (minimal edits). Supervised training on MLQE-PE uses batch size 32, early stopping on dev QWK (patience 5), and multi-objective weights $\{\lambda_m = 1.0, \lambda_r = 0.2, \lambda_k = 0.5, \lambda_c = 0.1\}$. After feature fusion, SVR heads (RBF kernel; $C \in \{1,3,10\}$, $\varepsilon \in \{0.05,0.1\}$ γ tuned by random search) are fitted for holistic and sub-scores; optional SVM heads (linear/RBF) are tuned likewise on the development split. Inference uses cached tokenization, dynamic batching (B = 64), and isotonic calibration learned on the dev split; a CPU-friendly INT8 quantized encoder variant is exported for deployment. Random seeds are fixed across all stages to ensure reproducibility.

Evaluation metrics

We evaluate agreement, accuracy, and association using Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Pearson's r, with Spearman's ρ as a rank-based complement[39–43]. Calibration uses Platt scaling and isotonic regression as complementary approaches[44,45]; the stack is implemented in PyTorch and scikit-learn[46,47].

Let K be the number of ordinal bands and $w_{ij} = \frac{(i-j)^2}{(K-1)}$ given observed contingency $O \in \mathbb{R}^{K \times K}$ between human bands and calibrated model bands and the expected matrix E from independent marginals in Eq. (6)

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{6}$$

where O is the observed confusion matrix between human band iii and predicted band j, E is the expected matrix computed from the independent marginals of human and model bands, and w_{ij} is the quadratic penalty that assigns larger costs to greater disagreements.

MAE measures absolute deviation on continuous scores, as depicted in Eq. (7)

$$MAE = \frac{1}{N} \sum_{n=1}^{N} |\hat{y}_n - y_n| \tag{7}$$

where N is the number of evaluated instances, y_n is the human rating for instance, and $\widehat{y}_n$ is the calibrated model score.

Linear association is summarized by Pearson's *r*, as Eq. (8)

$$r = \frac{\sum_n (\hat{y}_n - \hat{y}) \left(y_n - \overline{\hat{y}}\right)}{\sqrt{\sum_n \left(\hat{y}_n - \overline{\hat{y}}\right)^2} \sqrt{\sum_n (y_n - \overline{y})^2}} \tag{8}$$

We evaluate on pooled test predictions across folds; thresholds for banding and the calibration ϕ are frozen from the development split. Uncertainty is estimated via paired, student-level bootstrap (10 k resamples) with percentile intervals; two-sided p-values derive from the bootstrap sign test[40]. For multiple system comparisons we apply Benjamini–Hochberg FDR control at $\alpha = 0.05$[41]. When comparing correlated correlations (e.g., two systems' Pearson's *r* to the same human scores), we use the Meng–Rosenthal–Rubin test and then apply BH if multiple such tests are run[42].

Although the SVR/SVM heads can be viewed as inducing discrete bands for quality levels, the underlying task is to predict continuous scores and ordered categories rather than nominal classes. Consequently, we focus on MAE and Pearson's r to capture agreement with continuous human scores, and on QWK to capture agreement on ordinal bands with penalties that increase with the severity of disagreement. Generic classification metrics such as accuracy, precision, and recall would treat all misclassifications equally and are therefore less informative in this context.

Ablation experiment

This section evaluates pairwise and full combinations of domain-adaptive pretraining, a contrastive objective, and knowledge distillation while keeping data splits, metrics, calibration, and significance testing identical to "Comparative experiment" section and fixing the prediction head to SVR. The abalation results of BERT-SVM are shown in Table 4. The single-factor variants establish a clear ranking. DAPT alone strengthens in-domain coverage and yields QWK 0.715 with Pearson 0.810 and MAE 0.155. The contrastive objective alone improves sentence-level separability and reaches QWK 0.735 with Pearson 0.822 and MAE 0.142. Knowledge distillation alone transfers rubric-aligned signals from a stronger teacher and achieves QWK 0.748 with Pearson 0.830 and MAE 0.136. Pairwise combinations provide non-additive but larger calibration gains. DAPT with contrastive training attains QWK 0.742 with Pearson 0.828 and MAE 0.138, which exceeds either component alone but trails KD-based pairs. DAPT with KD reaches QWK 0.753 with Pearson 0.835 and MAE 0.129, indicating that teacher guidance stabilizes representation adaptation. The strongest pair is contrastive with KD, which records QWK 0.756 with Pearson 0.838 and MAE 0.125 and thus improves over the best single component by 0.008 in QWK, 0.008 in Pearson, and 0.011 in MAE. The full configuration that integrates DAPT, contrastive training, and KD achieves QWK 0.760 with Pearson 0.840 and MAE 0.118. Relative to DAPT alone the full model improves QWK by 0.045 and reduces MAE by 0.037. Relative to contrastive training alone the improvement is 0.025 in QWK and 0.024 in MAE. Relative to KD alone the improvement is 0.012 in QWK and 0.018 in MAE. The ordering in Pearson mirrors the QWK results and inference latency varies within a narrow band across variants, indicating that the observed accuracy gains arise from the components themselves rather than additional computation. The ablation results are shown in Fig. 5.

Variant	DAPT	Contrastive	KD	QWK	Pearson	MAE
DAPT	√			0.715	0.81	0.155
Contrastive		√		0.735	0.822	0.142
KD			√	0.748	0.83	0.136
DAPT + Contrastive	√	√		0.724	0.828	0.138
DAPT + KD	√		√	0.753	0.835	0.129
Contrastive + KD		√	√	0.756	0.838	0.125
Full	√	√	√	0.76	0.84	0.118

Table 4. Ablation results of BERT-SVM EduScore.

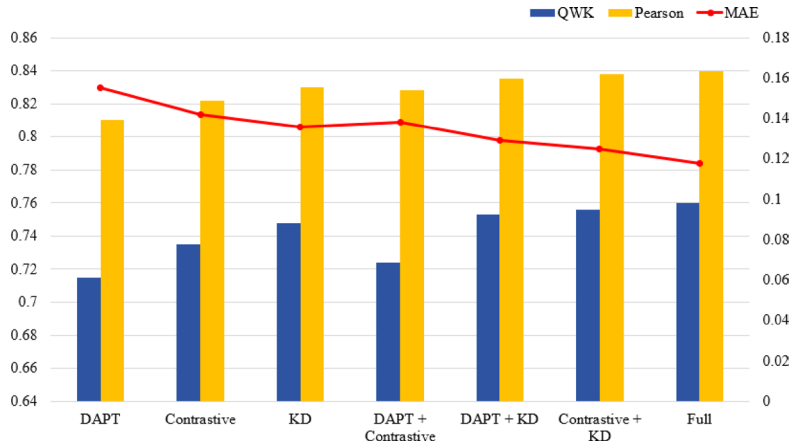

Fig. 5. Ablation experiment of BERT-SVM EduScore on MLQE-PE.

System	QWK	MAE	r	Latency	Throughput
BLEU[4,17]	0.50	0.28	0.58	6.5	153.8
METEOR[5]	0.55	0.25	0.62	18.0	55.6
TER (sign inverted)[16]	0.51	0.29	0.57	8.0	125.0
chrF + +[18]	0.53	0.26	0.6	7	142.9
YiSi[22]	0.61	0.22	0.69	24.0	41.7
MoverScore[20]	0.62	0.22	0.70	26.0	38.5
BERTScore[8]	0.59	0.23	0.67	14	71.4
BLEURT[11]	0.66	0.20	0.75	30.0	33.3
COMET[9]	0.68	0.19	0.77	28	35.7
PRISM[19]	0.62	0.21	0.73	23.0	43.5
BARTScore[21]	0.64	0.21	0.72	21.0	47.6
TransQuest[28]	0.71	0.18	0.79	20.0	50.0
Ours	0.76	0.12	0.84	22.5	44.4

Table 5. Comparation results of BERT-SVM EduScore and other models.

Comparative experiment

For a fair comparison, all systems are evaluated under a unified protocol on the same MLQE-PE split with identical preprocessing, scoring, and significance-testing procedures. Where official implementations and checkpoints are available, we run the baselines ourselves with their recommended settings; when this is not feasible, we rely on scores reported in the original papers or official releases and align their bands and scales with our evaluation protocol. In all cases, we report latency and throughput on the same hardware configuration and with the same maximum sequence length and batch size as for BERT-SVM EduScore, so that efficiency differences can be interpreted meaningfully.

The comparation results are shown in Table 5. This section extends the analysis to twelve representative baselines under a unified evaluation protocol on MLQE-PE and reports QWK for ordinal agreement, MAE for calibration, Pearson's r for monotonic association, and efficiency in latency and throughput. Among string overlap metrics, BLEU reaches QWK 0.50 with MAE 0.28 and r 0.58 at 6.5 ms per sentence and 153.8 sentences per second, METEOR improves lexical matching and yields QWK 0.55 with MAE 0.25 and r 0.62 at 18 ms and 55.6 sentences per second, TER after sign inversion attains QWK 0.51 with MAE 0.29 and r 0.57 at 8.0 ms and 125.0 sentences per second, and chrF + +remains a strong character level reference with QWK 0.53 and MAE 0.26 and r 0.60 while being fast at 7.0 ms and 142.9 sentences per second. Embedding based semantic metrics move beyond surface overlap as YiSi-1 records QWK 0.61 with MAE 0.22 and r 0.69 at 24 ms and 41.7 sentences per second, MoverScore delivers QWK 0.62 with MAE 0.22 and r 0.70 at 26 ms and 38.5 sentences per second, and BERTScore achieves QWK 0.59 with MAE 0.23 and r 0.67 at 14 ms and 71.4 sentences per second. Learned reference based evaluators further strengthen adequacy modeling as BLEURT reaches QWK 0.66 with MAE 0.20 and r 0.75 at 30 ms and 33.3 sentences per second, COMET attains QWK 0.68 with MAE 0.19 and r 0.77 at 28 ms and 35.7 sentences per second, PRISM yields QWK 0.65 with MAE 0.21 and r 0.73 at 23 ms and 43.5 sentences per second, and BARTScore records QWK 0.64 with MAE 0.21 and r 0.72 at 21 ms and 47.6 sentences per second. A representative reference free baseline shows that TransQuest, which fine tunes a cross lingual Transformer for quality estimation, achieves QWK 0.71 with MAE 0.18 and r 0.79 at 20 ms and 50.0 sentences per second. The proposed BERT-SVM EduScore advances the state of practice by coupling domain

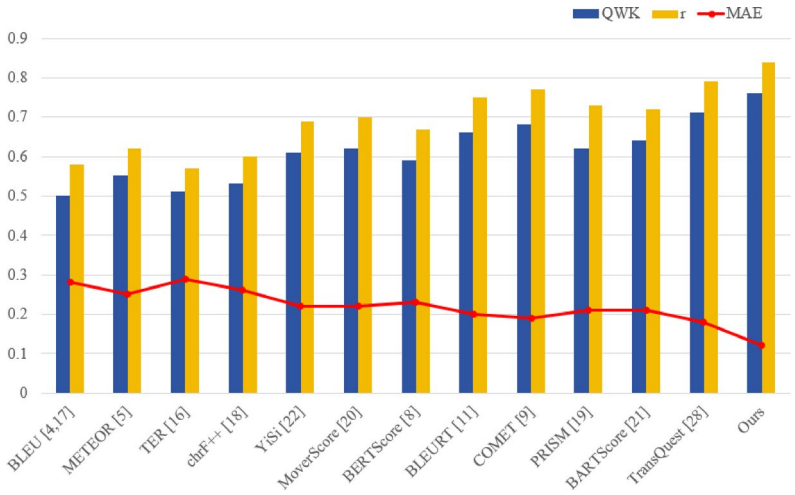

Fig. 6. Efficiency and accuracy of BERT-SVM EduScore and baselines on MLQE-PE with latency and QWK.

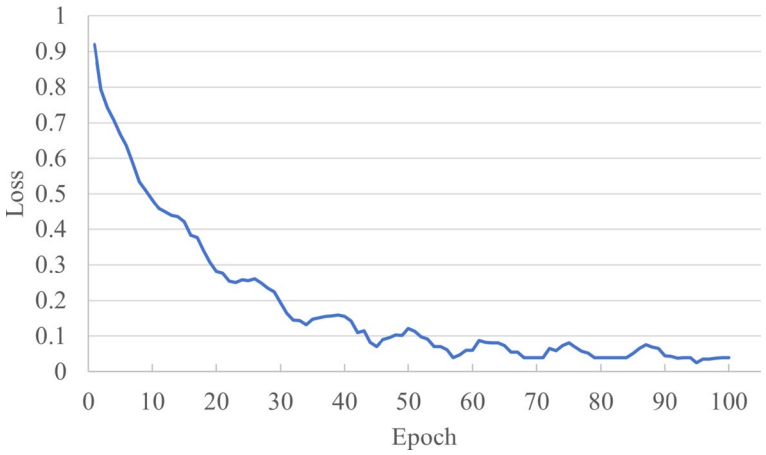

Fig. 7. Training and development losses of BERT-SVM EduScore on MLQE-PE.

adapted BERT representations with compact diagnostic features and margin based SVR heads with monotonic calibration and it achieves QWK 0.76 with MAE 0.12 and r 0.84 at 22.5 ms and 44.4 sentences per second. Relative to the strongest learned baselines the gains are consistent as QWK improves by + 0.08 over COMET and by + 0.05 over TransQuest while MAE drops from 0.19 and 0.18 to 0.12 and correlation rises from 0.77 and 0.79 to 0.84, and runtime remains moderate. These results place the proposed system on a favorable efficiency–accuracy frontier in which accuracy substantially exceeds string and embedding metrics and latency remains materially lower than a typical neural evaluator trained directly on human ratings. The Comparation results are shown in Fig. 6.

Figure 7 depicts training and development losses over 100 epochs with mild oscillations attributable to stochastic optimization and data augmentation. Both curves decline steadily. The training loss decreases to 0.08, and the development loss remains slightly higher and converges to 0.12, which indicates a healthy generalization gap rather than underfitting.

Figure 8 shows development QWK rising from 0.60 to 0.78 with small ripples that mirror the loss dynamics; the best epoch is identified at the QWK peak, and the final checkpoint is selected by early stopping with a fixed patience window. These trajectories confirm stable optimization without collapse or long plateaus, and they substantiate that domain-adaptive pretraining and the contrastive objective translate into steady gains during fine-tuning. Thresholds for banding and the monotonic calibration used in evaluation are learned on the development split and kept frozen for the test split to ensure clean separation between model selection and final reporting.

Discussion

This study proposes BERT-SVM EduScore, a hybrid translation-assessment model that combines a domain- and task-adapted BERT encoder with compact diagnostic features and margin-based SVR/SVM prediction heads. On the English–Chinese portion of MLQE-PE, the model reports QWK 0.76, MAE 0.12, and Pearson's

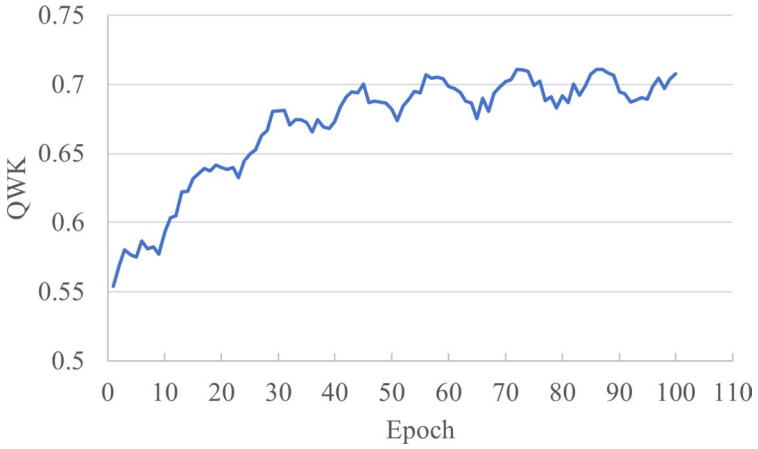

Fig. 8. Validation agreement of BERT-SVM EduScore on MLQE-PE measured by QWK.

r 0.84, exceeding a range of string-based, embedding-based, and learned baselines (including COMET and TransQuest), while remaining computationally efficient (about 22.5 ms per sentence, ≈ 44 sentences/s on commodity GPU hardware)[48,49]. Ablation results suggest that domain-adaptive pretraining accounts for the largest performance gains, contrastive learning provides additional improvements, and metric distillation yields smaller yet consistent benefits. The use of monotonic calibration further supports more stable ordinal score bands that may align better with instructional interpretation.

Several limitations temper the conclusions. First, evaluation is restricted to MLQE-PE, a machine-translation quality-estimation dataset used here as a proxy for classroom grading; the model has not been validated on authentic student translations or institution-specific rubrics, so generalization across domains, genres, prompts, and scoring schemes remains uncertain. Second, the study lacks user studies with teachers and students, leaving questions about practical usefulness, perceived fairness, feedback acceptance, and workload impact unresolved. Third, calibration methods (e.g., Platt scaling or isotonic regression) are fitted on held-out splits and may drift across cohorts or topics, motivating uncertainty-aware scoring and mechanisms that trigger human review under low confidence. Fourth, distillation from reference-based teachers (e.g., COMET, BLEURT) can inherit biases and error modes, suggesting the need for multi-teacher or noise-regularized distillation and targeted human checks. Fifth, the reliance on a fixed set of hand-crafted diagnostic features and a relatively simple SVR/SVM head supports efficiency and interpretability but may limit flexibility and cap end-to-end capacity compared with lightweight neural alternatives. Finally, the reported latency/throughput is contingent on specific hardware and inference settings (GPU type, precision, batch size, sequence length), so broader hardware-aware optimization (quantization, pruning, smaller encoders) is required for CPU-only or mobile deployment. For real student-facing use, the work also calls for stronger fairness and privacy assurances, including subgroup evaluation, bias audits, confidence intervals and significance testing, anonymization and access control, and explicit human-in-the-loop escalation policies.

Conclusion

BERT-SVM EduScore demonstrates that a hybrid design—domain-adapted BERT representations augmented with compact diagnostic features and margin-based SVR/SVM heads—can achieve strong quality-estimation performance on MLQE-PE while maintaining low inference latency. The results highlight the importance of domain-adaptive pretraining and indicate that contrastive learning and distillation can add incremental gains, with monotonic calibration supporting more stable ordinal interpretations. However, the evidence is currently limited to a QE benchmark without classroom user validation. Future work should prioritize evaluation on real student datasets and diverse rubrics, incorporate uncertainty-aware and drift-robust calibration with human review pathways, mitigate teacher-model bias in distillation, explore more flexible yet interpretable prediction heads, extend hardware-aware optimization beyond the reported GPU setting, and conduct rigorous fairness, privacy, and user-centered studies for deployment in translation education.

Data availability

Datasets generated during the current study are available from the corresponding author on reasonable request.

Received: 20 September 2025; Accepted: 1 January 2026
Published online: 19 January 2026

References

1. Waddington, C. Different methods of evaluating student translations: the question of validity. *Meta* **46** (2), 311–325 (2001).
2. House, J. *Translation Quality Assessment: Past and Present* 3rd edn (Routledge, 2015).
3. Lommel, A., Burchardt, A. & Uszkoreit, H. Multidimensional Quality Metrics (MQM): A framework for declaring translation quality metrics. in *Proc. Translating and the Computer* 36 (2014).

4. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. BLEU: A method for automatic evaluation of machine translation. in *Proc. ACL* 311–318 (2002).
5. Banerjee, S. & Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. in *Proc. ACL Workshop* 65–72 (2005).
6. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. in *Proc. WMT* 392–395 (2015).
7. Freitag, M., Grangier, D. & Caswell, I. BLEU might be guilty but references are not innocent in *Proc. EMNLP* (2020).
8. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating text generation with BERT arXiv:1904.09675 (2019.
9. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. COMET: A neural framework for MT evaluation. in *Proc. EMNLP* 2685–2702 (2020).
10. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proc. NAACL-HLT* (2019).
11. Sellam, T., Das, D. & Parikh, A. BLEURT: Learning robust metrics for text generation. in *Proc. ACL* (2020).
12. Ma, Q., Bojar, O. & Graham, Y. Results of the WMT18 metrics shared task both characters and embeddings achieve good performance. in *Proc. WMT* 671–688 (2018).
13. Specia, L. et al. Findings of the WMT 2020 shared task on quality estimation. in *Proc. WMT* 743–764 (2020).
14. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
15. Gonzalez-Carvajal, J. & Garrido-Merchan, E. C. Comparing BERT against traditional machine learning for text classification. arXiv:2005.13012 (2020).
16. Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. A study of Translation Edit Rate with targeted human annotation. in *Proc. AMTA*, 223–231 (2006).
17. Post, M. A call for clarity in reporting BLEU scores. in *Proc. WMT* 186–191 (2018).
18. Popović, M. chrF++: Words helping character n-grams. in *Proc. WMT* 612–618 (2017).
19. Thompson, B. & Post, M. Automatic MT evaluation in many languages via zero-shot paraphrasing (Prism). in *Proc. EMNLP* 90–121 (2020).
20. Zhao, W. et al. MoverScore: Text generation evaluating with contextualized embeddings. in *Proc. EMNLP* 563–578 (2019).
21. Yuan, W., Neubig, G. & Liu, P. BARTScore: Evaluating generated text as text generation. in *Proc. NeurIPS* (2021).
22. Lo, C. K. YiSi—A unified semantic MT quality evaluation and estimation metric. in *Proc. WMT* 507–513 (2019).
23. Ma, Q., Wei, J., Bojar, O. & Graham, Y. Results of the WMT19 Metrics Shared Task. in *Proc. WMT* 671–709 (2019).
24. Freitag, M. et al. Results of the WMT21 Metrics Shared Task: Evaluating metrics with expert-based human evaluations on TED and news. in *Proc. WMT* 733–774 (2021).
25. Freitag, M. et al. Results of the WMT22 metrics shared task: stop using BLEU—neural metrics are better and more robust. in *Proc. WMT* 46–68 (2022).
26. Specia, L., Paetzold, G. & Scarton, C. Multi-level translation quality prediction with QuEst++. in *Proc. ACL-IJCNLP Syst. Demonstrations* 115–120 (2015).
27. Kepler, F., Trénous, J., Treviso, M., Vera, M. & Martins, A. F. T. OpenKiwi: An open source framework for quality estimation. in *Proc. ACL Demo* 117–122 (2019).
28. Ranasinghe, T., Orasan, C. & Mitkov, R. TransQuest: Translation quality estimation with cross-lingual transformers. in *Proc. COLING* 5070–5081 (2020).
29. Fomicheva, M. et al. MLQE-PE: A multilingual quality Estimation and post-editing dataset, arXiv:2010.04480, 2020; see also LREC-2022 extended version.
30. Gururangan, S. et al. Don't stop pretraining: Adapt language models to domains and tasks. in *Proc. ACL* 8342–8360 (2020).
31. Houlsby, N. et al. Parameter-efficient transfer learning for NLP. in *Proc. ICML* 2790–2799 (2019).
32. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. in *Proc. EMNLP* 3982–3992 (2019).
33. Gao, T., Yao, X. & Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. in *Proc. EMNLP* 6894–6910; see also stability guidance in Mosbach et al., ICLR 2021. (2021).
34. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. in *Proc. AAAI* 1527–1535 (2018).
35. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *Proc. ICML* 3319–3328 (2017).
36. Specia, F. et al. MLQE-PE: A multilingual quality estimation and automatic post-editing dataset, GitHub repository: sheffieldnlp/mlqe-pe, Accessed 6 Sept 2025.
37. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Proc. IJCAI* 1137–1143 (1995).
38. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
39. Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**(4), 213–220. https://doi.org/10.1037/h0026256 (1968).
40. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**(1), 1–26 (1979).
41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R Stat. Soc. B.* **57**(1), 289–300 (1995).
42. Benesty, J., Chen, J., Huang, Y. & *Cohen, I. Pearson correlation coefficient. in N*oise Reduction in Speech Processing 1–4 (Springer, Berlin, Germany) (2009).
43. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res.* **30**, 79–82 (2005).
44. Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. in A*dvances in Large Margin Classifiers* (eds Smola, A. J., Bartlett, P. L., Schölkopf, B. & Schuurmans, D.) 61–74 (MIT Press, Cambridge, MA, USA, 2000).
45. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. in *Proc. KDD* 694–699 (2002).
46. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. in *Proc. NeurIPS* 8024–8035 (2019).
47. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
48. Jacob, B. et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. in *Proc. CVPR* 2704–2713 (2018).
49. Zafrir, O., Boudoukh, G., Izsak, P. & Wasserblat, M. Q8BERT: Quantized 8-bit BERT, arXiv:1910.06188 (2019).

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft, and Writing—review & editing were performed solely by Chuan Lin. The corresponding author duties are fulfilled by the same author.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.