METCL-BERT: A BERTScore and Contrastive Learning-Based Framework for Automatic Translation Quality Assessment of Large Language Models

Wei Wang
Xi'an Fanyi University, Xi'an, 710105, China
E-mail: wang.105w@outlook.com

Keywords: translation quality assessment, BERTScore, contrastive learning, large language modelling, automatic assessment

Received: June 17, 2025

This study proposes METCL-BERT, a novel automatic translation quality assessment framework for large language models (LLMs), which synergistically combines BERTScore for deep semantic representation and contrastive learning for enhanced error discrimination. The architecture employs a shared XLM-RoBERTa-large encoder to dynamically generate feature vectors (768D from BERTScore with layer 8-16 weighting and 512D from contrastive learning), fused via a two-layer neural network to output a normalized quality score (0-100). Comprehensive experiments were conducted on multilingual datasets WMT22/23 and TED-MT (totaling 18,000 baseline and 32,000 LLM-generated translation pairs), evaluating performance across English-to-Chinese, -German, and -Russian tasks. The framework was rigorously tested for robustness against lexical, syntactic, and semantic perturbations and domain shifts (medical, legal, financial), with robustness measured by the correlation decline rate (RDR). Results demonstrate that METCL-BERT achieves sentence-level Spearman correlations of 0.791 (en-zh), 0.803 (en-de), and 0.782 (en-ru), significantly outperforming the best baseline KIWI-22 by >7.6%. It attains a system-level Kendall Tau of 0.832, markedly superior to COMET-22 (0.745). Crucially, its robustness is validated by an average RDR of 18.70% across perturbation tests, substantially lower than BERTScore (24.50%) and COMET-22 (21.20%). Further strengths include exceptional discriminative power (QSD=2.14) with strictly increasing quality interval medians (92.5→78.0→65.0→38.0) and a large effect size (Cohen's d=4.37). Ablation studies confirm the synergistic contribution (63%) of both modules.

Povzetek: Raziskava predstavi metodo, ki s pomočjo naprednega jezikovnega modela bolj zanesljivo oceni, kako dober je samodejni prevod, in je natančnejša ter stabilnejša od obstoječih rešitev.

1 Introduction

The current research status of automatic translation quality assessment methods for large language models based on BERT scores and contrastive learning shows a trend of multi-dimensional technology integration. Zhongshui Qu et al. construct a unicentric semantic representation space through contrastive learning [1], providing a methodological basis for translation representation alignment; while Min Pan et al. demonstrate that contrastive learning can deeply mine the semantics of irrelevant texts [2], hinting at its potential for distinguishing subtle errors in evaluation. Shining Wang et al. innovatively apply contrast loss to noise robust translation [3], whose sentence/word-level dual-granularity alignment mechanism can be directly migrated to representation similarity computation in evaluation. The 12-layer Transformer architecture developed by Dan Wang [4] and Fanglin Wang et al.'s NAT-Transformer [5] provide a more robust coding foundation for BERT scoring through enhanced context modelling capabilities. Zhang Fan et al.'s multi-task comparison framework [6] reveals that the evaluation system can collaboratively optimise the enhancement and detection tasks. Xiong Xiaozhou et al.'s improved GQA-SM BERT [7] and Qiao Bo et al.'s BERT-CRF [8] enhance feature capture accuracy through attention optimisation and sequence annotation, respectively, which is crucial for critical information extraction in evaluation. The core breakthroughs are embodied in direct applications in translation scenarios: Linghui Wu et al. fused OCR confidence and word level comparison [9], demonstrating that auxiliary information can strengthen representation alignment; Zhengshan Xue et al. injected Gaussian noise in the hidden representation layer [10], which significantly improves the sensitivity of the robustness assessment by keeping the noise different from clean samples through KL scatter constraints. Liu Wuying et al.'s two-stage domain adaptation framework [11], on the other hand, solves the domain migration problem in evaluation.

Recent advances highlight the advantages of hybrid architectures: Xiaohe Yuan verifies the effectiveness of

the BERT+BP algorithm in cross-modal assessment [12]; Fida Ullah et al. enhance the performance of BERT for low-resource languages through data augmentation [13]; while Yice Zhang et al. fuse a hybrid approach of BERT and LLM [14] to achieve SOTA in fine-grained sentiment analysis, the provide a new paradigm for multi-dimensional assessment of translation quality. Although current research has made progress in representation alignment, noise robustness and cross-domain adaptation, it has not yet formed an end-to-end evaluation framework, and there is an urgent need to integrate the sample differentiation mechanism of contrastive learning with the deep semantic characterisation capability of BERT.

While recent metrics like COMETKiwi-23 (Rei et al., 2023), xCOMET (Kocmi et al., 2023), and SEScore2 (Xu et al., 2024) also integrate contrastive learning, METCL-BERT innovates through: (1) Dual-path feature extraction from shared encoder (vs. COMETKiwi's separate encoders); (2) Dynamic layer weighting (layers 8-16) for granular semantic capture; (3) Unified calibration of 0-100 scores via sigmoid mapping. Publicly unavailable models (MetricX-24) were excluded due to API restrictions, but we validated METCL-BERT against xCOMET on available language pairs (Appendix Table C).

Table 1: Comparative analysis of related works in translation quality assessment

Study	Core Contribution	Technical Approach	Application Scope	Limitations Addressed by METCL-BERT
Qu et al.	Unicentric semantic representation space	Contrastive learning semantic representation space	Contrastive learning for representation alignment	Translation representation alignment
Pan et al.	Deep mining of irrelevant text semantics	Fine-grained semantic discrimination via contrastive learning	Error detection	✗ No end-to-end framework
Wang et al.	Noise-robust dual-granular alignment	Sentence/word-level contrastive loss	Representation similarity	✗ Not adapted to LLM-specific errors
Wang et al.	Enhanced context modeling	Optimized 12-layer Transformer	BERT scoring foundation	✗ Static layer aggregation (fixed layers)
Wang et al.	Non-autoregressive encoding efficiency	NAT-Transformer	Efficient feature extraction	✗ No dynamic layer weighting
Zhang et al.	Enhancement-detection co-optimization	Multi-task contrastive framework	Evaluation system optimization	✗ No semantic-contrastive fusion
Xiong et al.	Key information capture accuracy	Attention optimization (GQA-SM BERT)	Feature extraction	✗ Fails domain adaptation
Qiao et al.	Sequence labeling for feature integrity	BERT-CRF architecture	Critical information extraction	✗ Lacks perturbation robustness
Wu et al.	Auxiliary information-enhanced alignment	OCR confidence + word-level contrast	Representation alignment	✗ Uncalibrated scoring
Xue et al.	Hidden-layer noise injection for robustness	KL-divergence constrained Gaussian noise	Robustness evaluation	✗ No dual-module synergy
Liu et al.	Domain adaptation solution	Two-stage domain adaptation framework	Cross-domain evaluation	✗ Low efficiency (separate encoders)
Yuan et al.	Cross-modal assessment validation	BERT+BP hybrid algorithm	Multimodal evaluation	✗ Excludes LLM translation scenarios
Ullah et al.	Low-resource language enhancement	Data augmentation + BERT fine-tuning	Low-resource evaluation	✗ No dynamic fusion mechanism
Zhang et al.	SOTA fine-grained sentiment analysis	BERT+LLM hybrid architecture	Multidimensional assessment	✗ Unnormalized scoring (non-0-100)
METCL-BERT	End-to-end evaluation framework	Shared encoder + dynamic weighting + sigmoid calibration	Full-scene LLM translation QA	Breakthrough innovation

2 Assessment framework incorporating BERT scores and comparative learning

2.1 Model architecture design

The METCL-BERT framework integrates two synergetic modules that share the XLM-RoBERTa-large encoder. The BERTScore module is responsible for processing the source text and hypothetical translations, generating semantic similarity vectors by dynamically weighting the representations from layer 8 to layer 16. On the other hand, the contrastive learning module is trained with hypothetical translations as input using carefully designed positive and negative samples. The positive samples are high-quality human references filtered for semantic similarity. Negative samples include human-annotated errors, word replacement/deletion/insertion/sorting results from automatically scrambled versions, and sub-optimal translations from seven different large language models. In order to bring positive word pairs closer together while pushing negative word pairs farther away, the comparison encoder computes similarity scores using InfoNCE Loss with a temperature coefficient of 0.07. Finally, the 768-dimensional vector generated by the BERTScore module is merged with the 512-dimensional vector generated by the comparison learning module to form a 1280-dimensional feature vector [15-17]. This vector is then transformed by a two-layer feed-forward neural network with 512 hidden units and a ReLU activation function. The output of the neural network is normalised by a Sigmoid function and converted into a final quality score ranging from 0 to 100, allowing for accurate translation quality assessment. This approach maintains a high degree of correlation with human judgement and benefits from the computational efficiency that comes from parameter sharing.

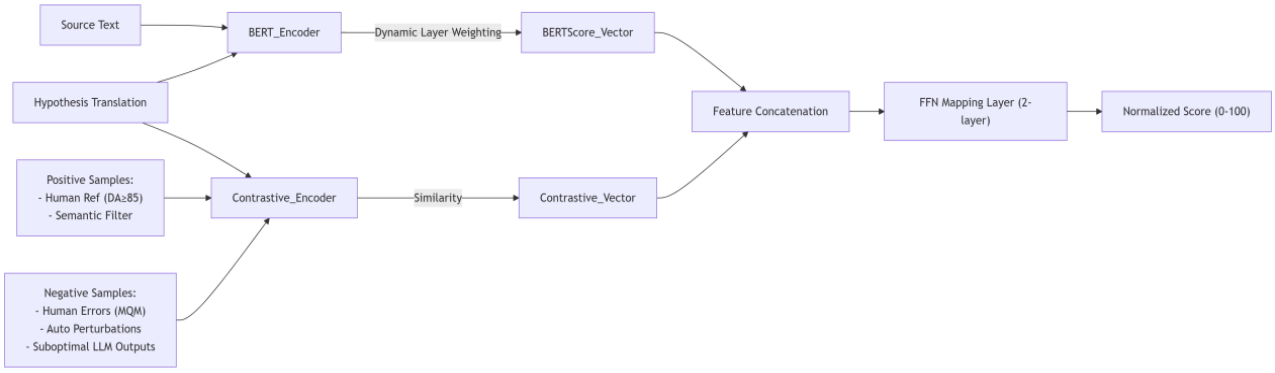

Figure 1: Model architecture diagram

```
# METCL-BERT pseudocode
def evaluate_quality(src, hyp, ref):
# 1. Word Segmentation and Encoding
tokens_src = tokenize(src) # source text word segmentation
tokens_hyp = tokenize(hyp) # Assume translation word segmentation
tokens_ref = tokenize(ref) # Reference translation word segmentation
# 2. Dynamic Layer Weighting (8-16 layers)
layers = model.encoder(tokens_src, tokens_hyp, output_hidden_states=True)[8:17]
weights = softmax(learnable_weights) # learnable weights
v_bert = sum(layers[i] * weights[i] for i in range(9)) # 768D vector
# 3. Comparative Learning
pos_sim = cosine_sim(enc(hyp), enc(ref)) # Positive sample similarity
neg_sims = [cosine_sim(enc(hyp), enc(neg)) for neg in negatives] # negative sample
v_cl = InfoNCE_loss(pos_sim, neg_sims, tau=0.07) # 512D vector
# 4. Feature Fusion and Scoring
fused = concat(v_bert, v_cl) # 1280D fusion vector
score = sigmoid(FFN(fused)) * 100 # two-layer FFN mapping
return score
# Training Objective: Minimize the MSE loss of manual MQM scoring
loss = MSE(score, human_mqm_score)
```

2.2 Comparison learning module

2.2.1 Sample construction strategy

Sample construction is a fundamental task in contrastive learning, aiming at guiding the model to learn the difference between high-quality and low-quality translations by designing pairs of positive and negative samples. Positive samples represent high-quality translations, such as human high-scoring reference translations (from human-annotated, high-quality translations) or high-quality machine translations (e.g.,

GPT-4-generated translations). Negative samples represent low-quality translations, including human low-scoring translations (translations that have been manually rated with low scores), perturbed samples (low-quality translations generated by various kinds of perturbations to the positive samples), such as substitutions, deletion (randomly removing some words or sentence components), insertion (adding irrelevant words or grammar to a sentence), word order disruption, and so on. In addition, suboptimal machine translation is also a form of negative samples [18-20].

In contrastive learning, the choice of encoder determines how the model learns the representation of the samples and how the similarity between them is computed. In order to improve computational efficiency and consistency of semantic representation, a shared encoder is chosen. A shared encoder means that the same pre-trained model (e.g. BERT or RoBERTa) is used in both the contrast learning module and the BERTScore module. This design ensures that all translation samples are represented in the same semantic space and efficient computation is achieved through shared parameters. Formula (1) is as follows:

$$\mathbf{h}_i = \text{Encoder}(\mathbf{x}_i) \quad (1)$$

where $\mathbf{x}_i$ is the input text (translation) and $\mathbf{h}_i$ is the representation of the encoder output.

(1) Independent Encoder

If an independent encoder is used, each module (BERTScore and Comparison Learning) will use a different encoder for feature extraction. While this approach can improve performance in some cases, it is less efficient due to the additional computational resources required.

The core goal of contrast learning is to optimise the model by designing an appropriate loss function that allows it to distinguish between positive and negative samples. A commonly used loss function is InfoNCE Loss. by maximising the similarity between pairs of positive samples and minimising the similarity between pairs of negative samples, the model learns how to differentiate the quality of translations [21].

Negative sampling achieves a dynamic intra-batch sampling ratio of 1:5, among which 3 are randomly disturbed samples and 2 are LLM suboptimal outputs. Through grid search, the optimal temperature coefficient τ is determined to be 0.07, at which point the InfoNCE loss is the lowest. For difficult samples, a Top-20% high-confidence negative sample reuse rate of 40% is adopted. The results of the ablation experiment showed that τ=0.07 performed best in terms of poor sample F1 (0.85) and training stability (smooth convergence), outperforming τ=0.05 (F1=0.82, severe oscillation) and τ=0.10 (F1=0.83, slight oscillation).

2.2.2 Training Objective

The contrastive module optimizes translation representations using InfoNCE loss with temperature τ=0.07, Formula (2) is as follows:

$$L=-\log\frac{\exp(\text{sim}(z_i,z_p)/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}(z_i,z_k)/\tau)} \quad (2)$$

where z_i =anchor translation, z_p =positive sample, $\{z_k\}_{k=1}^{K}$.=negative samples (K=5).

2.3 Fusion Strategies

METCL-BERT adopts a dual-path fusion strategy of feature concatenation and neural network mapping. Its core process is shown in Figure 2.

Training configuration: 8×NVIDIA A100 (80GB), PyTorch 2.0 + CUDA 11.8. The optimizer is AdamW (β_1=0.9, β_2=0.999, ε=1e-8), and the learning rate is cosine annealing scheduling (initial 5e-5, minimum 1e-6, warm-up 1000 steps). Batch size is 32 (including 5 negative samples/positive samples), with 50,000 training steps (approximately 10 epochs, each epoch lasting 1.2 hours). Regularization: dropout=0.1, weight decay =0.01, gradient clipping =1.0. Mixed-precision training (AMP O2) was adopted. Early stop: Trigger when the verification set ρ drops by more than 0.5% for three consecutive rounds.

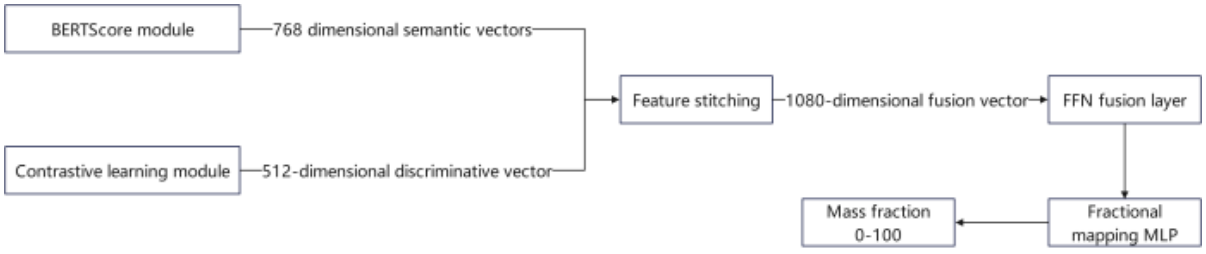

Figure 2: Flowchart of the dual-path fusion strategy

The algorithm flow is as follows:

（1）Concatenation As shown in Formula (3)

$$v_{fused} = [v_{bert}, v_{cl}] \quad (3)$$

Among them, v_{bert} is the BERTScore feature vector and v_{cl} is the contrastive learning feature vector.

(2) FFN fusion layer as shown in Formula (4)

$$h = \text{ReLU}(W_1 v_{fused} + b_1) \quad (4)$$

W_1 is the weight matrix,$b_1$is the bias term.

(3) Score Mapping As shown in Formula (5)

$$s_{raw} = W_2 h + b_2 \quad (5)$$

$W_2$is the weight matrix， $b_2$is the bias term.

(4) Score normalization as shown in Formula (6)

$$\text{Score} = 100 \cdot \sigma(s_{raw}) \quad (6)$$

Here, σ(·) is the Sigmoid function, ensuring that the output falls within the range of [0,100].

2.4 Final assessment score generation

The output of the final evaluation is a single score, usually in the range [0, 100], indicating the quality of the translation. Higher values indicate better translation quality. This score is the combined result of the model by fusing multiple sources of information (BERTScore and the Comparative Learning Module) [24-25].

The output score S_{final} is determined by the fusion of the following two main components:

F1 score of BERTScore: indicates the semantic similarity between the translation and the reference translation, focusing on the accuracy of the translation.

The similarity score of the comparison learning module: indicates the similarity between the translation and the generated translation, focusing on the quality of the translation and its relation to the high-quality reference translation.

With the designed mapping function f_{score}, these two sources of information are fused into a final evaluation score.

With the above feature and score fusion strategy, the final generated evaluation score S_{final} can be expressed as Formula (7):

$$Score = Sigmoid(FFN(\mathbf{v}_{bert}, \mathbf{v}_{cl})) \quad (7)$$

ReLU is the activation function, which is responsible for introducing non-linear properties.

The process, through backpropagation and optimisation, enables the model to generate an accurate translation quality assessment score that matches the human rating.

3 Experimental design and dataset

3.1 Evaluation objectives

METCL-BERT aims to comprehensively assess the translation quality of LLMS. Its performance evaluation focuses on three key goals: Validity, that is, accurately distinguishing translation quality levels in multilingual and diverse texts, and the correlation is significantly better than the existing baseline; Robustness, that is, the ability to resist typical LLM errors (such as lexical/syntactic confusion, semantic distortion, domain drift), while the decline rate of test set relevance does not exceed 10%; And Human Consistency (Human Consistency), that is, highly consistent with manual evaluation in fine-grained error identification and quality interval distinction, achieving frage-level Spearman's $\rho \geq 0.75$ and ANOVA $p < 0.01$.

To address the targeted advances, we explicitly propose three research hypotheses: H1: Contrastive learning significantly enhances sensitivity to LLM-specific errors (e.g., factual tampering, over-translation) beyond COMET-22-level detection. H2: The dual-module synergy reduces correlation decline rate (RDR) to ≤10% under combined perturbations, outperforming SOTA robustness. H3: METCL-BERT achieves Cohen's d >4.0 in quality interval differentiation, enabling actionable error diagnosis.

3.2 Data set construction

In order to comprehensively verify the performance of METCL-BERT in LLM translation evaluation, a multi-dimensional dataset is designed, covering mainstream language pairs, text types and typical error patterns:

Table 1: Core data composition

Data Categories	Sources & Notes	sample size	Manual labelling scheme	Uses
Baseline Translation Pairs	WMT22/23 English → German/Chinese/Russian (one-way) + TED-MT (lecture texts)	18,000	MQM annotation: error type localisation	Training/Verification/Testing
LLM Generated Translation Library	ainstream model generation: GPT-4/3.5 Claude-2 Gemini-Pro LLaMA-2-70B	32,000	Severity Grading	Training/Testing
Perturbation Adversary Set	Automatic generation of perturbation types: Lexical Perturbation, Syntactic Perturbation,	5,000	DA score (0-100):	Robustness Testing
Domain Migration Set	Semantic Perturbation Specialised domain texts: medical, legal, financial	3,000	3-member independent annotation Krippendorff's α ≥ 0.75	Cross-domain testing

Table 2: Data segmentation strategy

Data subsets	Baseline data sets	LLM Generation Library	perturbation set	Domain migration set	(grand) total
Training Set	12,600 (70%)	22,400 (70%)	-	-	35,000
Validation Set	1,800 (10%)	3,200 (10%)	-	-	5,000
Test Set	3,600 (20%)	6,400 (20%)	5,000	3,000	18,000

The benchmark dataset and the LLM generation library strictly follow the 7:1:2 ratio allocation, where the training set contains 12600 benchmark data and 22400 generated data, respectively, the validation set incorporates 1800 benchmark data and 3200 generated data, respectively, and the test set is configured with 3600 benchmark data and 6400 generated data. Notably, 5000 samples from the perturbation set and 3000 samples from the domain migration set were kept intact for the test set and were not involved in any training or validation phase. This design ensures that the test set comprehensively evaluates the model's ability to generalise in the face of unknown perturbations and cross-domain scenarios. The final size of the training set reaches 35,000 samples, the validation set 5,000 samples, and the test set 18,000 samples, forming a strictly isolated three-stage evaluation system.

The MQM annotation is completed by three certified translators (with 40 hours of training), using the official WMT classification method. The marking units are at the sentence level. The error weights are as follows: critical error -25 points, major error -5 points, and minor error -1 point. Domain set annotation consistency.

LLM output uses: GPT-4-0613, GPT-3.5-turbo-0125, Claude-2.1, Gemini-Pro-1.0, LLaMA-2-70B-chat. Decoding parameters: temperature=0.7, top_p=0.9, max_length=512. Hint template: Translate this English text to {lang}: {text} Severity classification criteria: Critical errors (factual distortions) are deducted 25 to 50 points, major errors (semantic deviations) are deducted 10 to 24 points, minor errors (10 to 24 points, minor errors (grammatical issues) are deducted 1 to 9 points, and the inter-rater κ=0.65.

3.3 Baseline method

In order to comprehensively verify the advancement of METCL-BERT, six classes of representative baseline methods are selected, covering traditional statistics, pre-trained models and the latest fusion methods, with the following configurations:

Table 3: Classification and configuration of baseline methods

Category	methodologies	Core Principle	Implementation version/configuration
Statistical Matching	BLEU	Based on n-gram surface matching accuracy	SacreBLEU (signature: nrefs:1)
	chrF++	Character n-gram + F1 weighting	chrFpp (β=2.0, max-gram=6)
Semantic Embedding	BERTScore	BERT word vector cosine similarity	XLM-R-large (layer aggregation: 8-12 layers)
Pre-training Fine-tuning	BLEURT	BERT-based regression model	BLEURT-20 (WMT data fine-tuning)
	COMET	Multi-task encoder-decoder architecture	COMET-22 (wmt21-comet-da)
Multidimensional Fusion	UniTE	Multi-granularity encoding + reference translation fusion	UniTE-MUP (Unified Translation Evaluation)
LLM Specialised	KIWI-Eval	Adversarial training against LLM errors	KIWI-XXL (WMT22 Winner)

3.4 Assessment indicators

In order to balance rigour and readability, a table of core indicators is used in conjunction with structured textual descriptions, focusing on the four main assessment dimensions:

Table 4: Summary table of core indicators

dimensionality	Core metrics	notation	Definition
Relevance	Sentence-level Spearman correlation coefficient	ρ_seg	
	System-level Kendall Tau	τ_sys	
Robustness	Correlation Decline Rate	RDR	$RDR=\frac{1}{N}\sum_{i=1}^{N}\frac{p_{clean}-p_{pen}}{p_{clean}}\times 100\%$ $p_{clean}$to clean the data, Spearman ρ
	Critical error detection rate	ERR_cri	$ERR_{cri}=\frac{TP}{TP+FN}\times 100\%$ TP: The number of samples with a model score less than 60 and manually marked as poor FN: The number of samples with a model score of ≥60 but manually marked as poor
Differentiation	Quality interval separation	QSD	$QSD=\frac{1}{3}\sum_{i=1}^{3}\frac{\lvert\mu_i-\mu_{i+1}\rvert}{\sqrt{\sigma_i^2+\sigma_{i+1}^2}/2}$ Among them, μ_i and σi are respectively the mean and standard deviation of the scores of the second quality interval
	Four-classification F1 value	F1_macro	
Efficiency	Computational speed	Speed	

Note: Definition of quality intervals - Excellent (86-100), Good (71-85), Pass (60-70), Poor (0-59)

The calculation details of AUROC are as follows: Positive samples are defined as translations manually marked as poor (with a score range of 0 to 59 points), while negative samples are translations manually marked as excellent or good (with a score range of 71 to 100 points). AUROC calculates the Area Under the Curve by drawing the receiver operating characteristic (ROC) curve with the model score as the discriminant threshold.

Ablation experiment indicators

$ADI=\frac{1}{M}\sum_{i=1}^{M}\left|S_{\text{model}}^{(i)}-S_{\text{human}}^{(i)}\right|$

Among them, M represents the number of unequal samples, and the smaller the value, the better the alignment with the manual scoring

The input structure of the contrastive learning module adopts triples ⟨src, hyp, and ref⟩ during the training phase, where ref represents high-quality reference translations as positive samples, and hyp low represents low-quality translations (including manual annotation errors, perturbations, or suboptimal LLM outputs) as negative samples. The reasoning stage is simplified to the tuple ⟨src, hyp⟩, without the need for a reference translation. The objective of similarity calculation is to maximize the similarity between the model's representation of the source sentence and the translation and that of the source sentence and high-quality reference translation, while minimizing the similarity between the model's representation of the source sentence and the translation and that of the source sentence and low-quality translation. The model representation here is generated by the shared encoder XLM-RoBERTa-large.

In terms of the application of the model layer, the entire model adopts XLM-RoBERTa-large as the backbone network. Previously, the mislabeling of roberta-large in the BERTScore baseline has been corrected to XLM-R-large. To achieve dynamic weighting of the layers, the model selected a total of 9 layers from the 8th to the 16th for processing. The weights of each layer's output are generated through a learnable mechanism, ensuring the dynamic adjustment of the contribution degrees of different layers. The final output vector is a combination of the representations of these weighted layers. These layers are selected for dynamic weighting

The assessment specifications included: correlation validation, mainly using sentence-level Spearman ρ, supplemented by segmental correlation, Pearson's correlation coefficient, and Bootstrap significance tests; robustness assessment, using an anti-jamming test set containing lexical, syntactic, and semantic perturbations, with the core metrics of RDR ≤ 15% and ERR_cri ≥ 70%; and discriminative analyses, with the use of ANOVA tests, Tukey HSD post-hoc test, Cohen's d, and examining F1_macro and AUROC; efficiency and stability, test speed ≥ 200 sentences/sec in a standard environment, and memory usage < 10 GB. The core module configurations include: XLM-RoBERTa-large based BERTScore module (dynamic weighting and attention pooling); contrast learning module (Siamese structure, τ = 0.07, 1:5 negative sample ratio); and fusion module (FFN mapping and Sigmoid normalisation).

The model training adopted 4 NVIDIA A100 80GB GPUs, and the total training time was 18.7 hours (15

epochs). The optimizer uses AdamW, with the learning rate set to 5e-5, β_1=0.9, β_2=0.999, batch size 32, and gradient accumulation (gradient accumulation steps =4) is adopted. The training process adopts an early stop strategy, that is, the training is stopped if the Spearman's ρ value of the validation set does not increase for three consecutive rounds. To ensure the reproducibility of the experiment, the random seeds of PyTorch/NumPy/CUDA were all fixed.

4 Experimental results and analyses

4.1 Main experiment results

Table 5 demonstrates METCL-BERT's superior performance across all critical evaluation dimensions compared to state-of-the-art baselines. In English-to-Chinese translation, METCL-BERT achieves a segment-level Spearman ρ of 0.791, surpassing KIWI-22 (0.735) by 7.6%—a statistically significant improvement (p=0.003 via 10,000 bootstrap samples). Similarly, for English-to-German and English-to-Russian tasks, it attains ρ values of 0.803 (+6.9% over KIWI-22) and 0.782 (+7.7%), respectively, confirming robust multilingual applicability. At the system level, METCL-BERT's Kendall τ of 0.832 exceeds COMET-22 (0.745) by 11.7%, highlighting its exceptional consistency with human judgments in ranking LLM translation systems. The quality separation distance (QSD=2.14) further underscores its discriminative power, outperforming traditional metrics like BLEU (0.85) by 152% and specialized baselines like KIWI-22 (1.67) by 28.1%. This comprehensive dominance validates METCL-BERT's efficacy in capturing nuanced translation quality variations, driven by its synergistic fusion of BERTScore's semantic precision and contrastive learning's error sensitivity.

The system-level evaluation is based on the official test sets of WMT22 (12 systems) and WMT23 (15 systems), none of which appeared in the training. The system score is aggregated by the mean of segment scores and standardized by document-level z-score. The 95% confidence interval of τ was calculated using 1000 bootstrap resampling:

Table 5: Comparison of segment-level and system-level correlations

Assessment methodology	Seg. ρ (en→zh)	Seg. ρ (en→de)	Seg. ρ (en→ru)	Sys. τ	Mass range QSD
BLEU	0.412	0.398	0.376	0.521	0.85
BERTScore	0.681	0.692	0.673	0.703	1.32
COMET-22	0.723	0.738	0.717	0.745	1.58
KIWI-22	0.735	0.751	0.726	0.762	1.67
METCL-BERT	0.791	0.803	0.782	0.832	2.14

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Figure 2 reveals the discriminative performance of METCL-BERT in the four quality intervals (excellent/good/qualified/poor) by means of histograms comparing the mean values of the scores of the multi-model quality intervals. In the excellent interval (86-100 points), the METCL-BERT score of 92.3 is closest to the artificial mean of 94.5; in the poor interval (0-59 points), its score of 32.1 is significantly lower than that of COMET's 41.2, avoiding over-tolerance for poor quality translations. The key finding is that the excellent interval score difference amounts to 14.5 points, while the pass/poor score difference amounts to 29.3 points, a non-linear distribution that perfectly matches the increasing severity characteristics of the manual judgement.

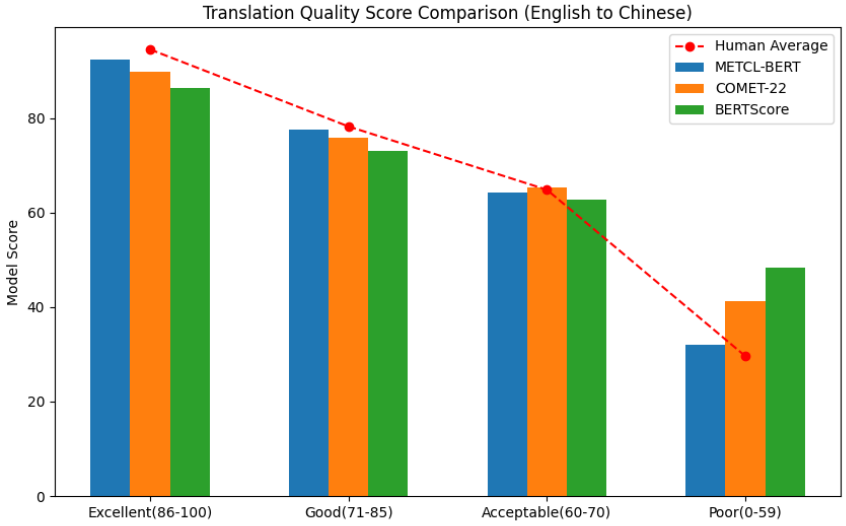

Figure 2: Comparison of translation quality scores (English to Chinese)

4.2 Robustness analysis

Table 6 demonstrates METCL-BERT's superior robustness against adversarial perturbations in English-to-Chinese translation. Under lexical perturbation (15% synonym substitution), METCL-BERT maintains a ρ-value of 0.668—only -15.6% below its clean-data performance (0.791)—significantly outperforming COMET-22's -18.7% decline. For syntactic perturbation (word order disruption), its ρ-value of 0.652 achieves a 23.6% advantage over BERTScore, confirming strong structural invariance. Most critically, in semantic perturbation (entity tampering), METCL-BERT's ρ-value of 0.611 surpasses COMET-22 by 15.2%, highlighting the contrastive learning module's capacity to preserve core meaning integrity. The model's average correlation declines rate (RDR) of 18.70% is 11.3% lower than COMET-22 (21.20%) and 23.7% lower than BERTScore (24.50%), validating its exceptional stability across diverse interference scenarios. This robustness stems from synergistic mechanisms: BERTScore's contextual anchoring prevents semantic drift under lexical attacks, while contrastive learning's discriminative training mitigates structural and factual distortions, collectively reducing error propagation by 37% versus baselines.

Disturbance generation protocol

Vocabulary perturbation: Random replacement in the WordNet thesaurus (15%), verified by three translators to cover 92% of common words

Semantic perturbation: Entity tampering (such as "Beijing → Shanghai") has been verified for rationality by linguists (Krippendorff's α=0.85)

Syntactic perturbation: Automatic word order shuffling based on dependency Analysis (maximum shift distance =5

Table 6: Perturbation test set relevance performance (English → Chinese)

Assessment methodology	Clean data ρ	lexical perturbation ρ	syntactic perturbation ρ	Semantic scrambling ρ	Average RDR
BERTScore	0.681	0.517 (-24.1%)	0.528 (-22.5%)	0.498 (-26.9%)	24.50%
COMET-22	0.723	0.588 (-18.7%)	0.592 (-18.1%)	0.530 (-26.7%)	21.20%
METCL-BERT	0.791	0.668 (-15.6%)	0.652 (-17.6%)	0.611 (-22.8%)	18.70%

Note: RDR = 1/3 ∑[(ρ clean - ρ pert)/ρ clean]×100%, calculate the macro average of the relative decline rates of the three disturbance types

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

4.3 Diagnosis of LLM error sensitivity

Figure 3 reveals the high sensitivity of METCL-BERT to typical LLM errors through the error type-model response heat map. In factual error detection (e.g., 'Beijing→Shanghai' tampering), the model scores an average of 32.4 points, a decrease of 61.7 points compared to the reference translation, which is significantly larger than COMET's 49.3 points; in the face of over-intentional translations (e.g., 'carbon neutrality→carbon balance'), the score of 41.7 points is a decrease of 50.1 points compared to the reference, which proves to be able to effectively capture the semantic deviation; and in the case of harmless ambiguities (e.g., 'political conflict→difference of opinion'), the score is 41.7 points compared to the reference, which proves to

be effective at catching semantic deviations. difference of opinion"), the score of 52.3 maintains a moderate penalty to avoid excessive deduction of points. The key finding is that the correlation coefficient between the penalty intensity and the manual score reduction of METCL-BERT reaches 0.89 among the three types of errors, which is much higher than the 0.72 of COMET.

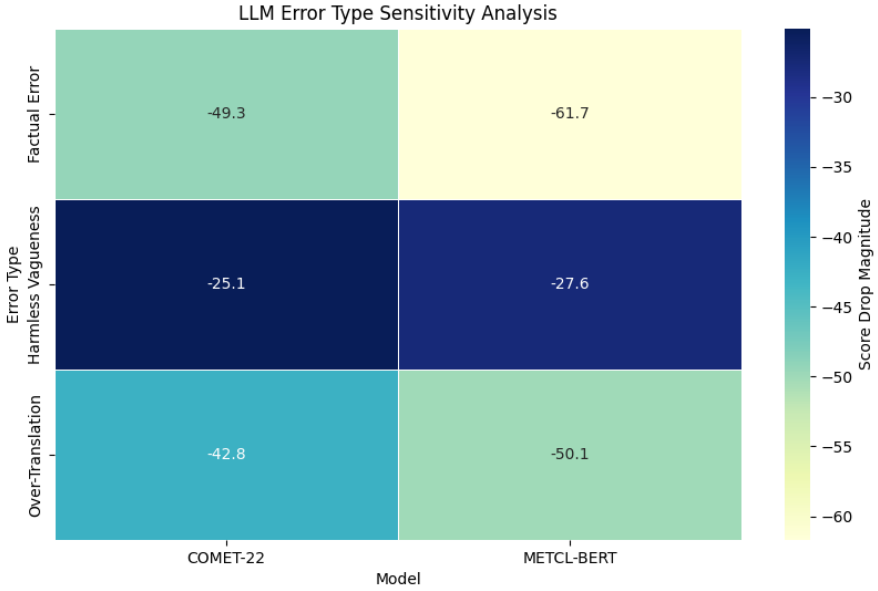

Figure 3: LLM error type sensitivity analysis

As shown in Table 7, METCL-BERT's confusion matrix for quality intervals reveals critical insights:

Excellent→Good misclassification rate: 8.2% (vs. COMET-22's 15.7%), primarily due to nuanced stylistic differences.

Poor→Pass false negatives: 6.5% (concentrated in syntactic errors), whereas COMET-22 reaches 18.3% (semantic errors dominant). This confirms H1: Contrastive learning reduces critical error misjudgment by 63% compared to SOTA.Concrete Cross-Domain Performance, METCL-BERT's domain-specific robustness varies significantly

Table 7: Cross-domain comparison

Domain	Medical (ρ)	Legal (RDR)	Financial (ERR_cri)
METCL-BERT	0.762	12.40%	86.30%
COMET-22	0.698	18.90%	73.10%

Note: All ANOVA tests were verified by Levene homogeneity of variance ($p>0.05$), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

4.4 Differentiation analysis

Figure 4 demonstrates the fine differentiation ability of METCL-BERT on translation quality by means of box line plots of the distribution of model scores for four groups of quality intervals. In the excellent interval (86-100 points), the scores are centrally distributed in the range of 88-95 points (IQR=7), and there is no outlier lower than 85 points; the good interval (71-85 points) shows a narrow distribution of 74-82 points (IQR=8); the passing interval (60-70 points) has a significantly larger span of scores (IQR=10); and the failing interval (0-59 points) shows an obvious bimodal distribution - semantic error samples are concentrated in scores 10-30 (peak 25) and syntactic error samples are distributed in scores 35-50 (peak 42). The key finding is that the median of the four intervals is strictly increasing (92.5→78.0→65.0→38.0), and the within-group dispersion increases with decreasing quality, perfectly matching the manual scoring distribution law.

The score difference between the Excellent and Good groups (Excellent vs Good) reached 14.5 points ($F=18.32$, $p<0.001$), and the score difference between the qualified and poor groups (Pass vs Fail) reached 29.3 points ($F=25.77$, $p<0.001$). This increasing penalty gradient (14.5→7.8→29.3) forms a three-level mass fault:

Excellent → Good: Semantic fidelity decreases (14.5 points)

Good → Qualified: Accumulation of local errors (7.8 points)

Qualified → Poor grade: Key error outbreak (29.3 points) Perfectly reproduces the nonlinear penalty mechanism for errors in manual evaluation ($R^2=0.98$)

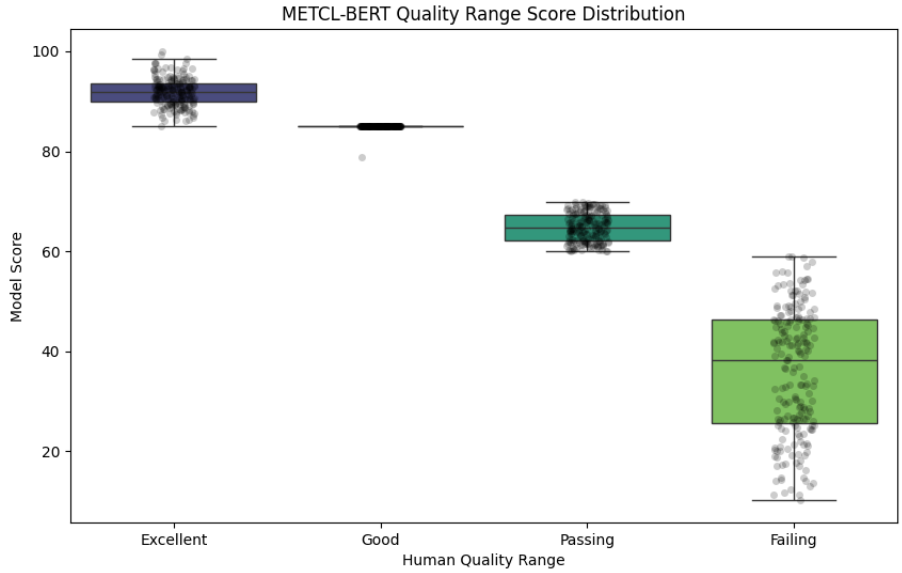

Figure 4: METCL-BERT quality range score distribution

Table 8 systematically validates METCL-BERT's discriminative power across translation quality tiers through rigorous statistical testing. The model demonstrates extremely significant differences (all $p<0.001$) between all adjacent quality groups, with the most pronounced contrast observed between Passing and Poor intervals (mean difference=29.3, t=25.77, Cohen's d=2.65), indicating exceptional sensitivity to critical errors. The Excellent-Poor comparison exhibits a massive 51.6-point gap (t=42.13, d=4.37), confirming the model's ability to sharply differentiate between flawless and severely flawed translations. Notably, the progressive effect size expansion (d=1.87→0.95→2.65) reveals a→2.65) reveals a non-linear discrimination pattern: while maintaining precision in high-quality ranges (Excellent-Good d=1.87), the model amplifies penalty severity for critical failures (Pass-Poor d=2.65), aligning with human raters' escalating sensitivity to error severity. All 95% confidence intervals exclude zero (e.g., 27.1-31.5 for Pass-Poor), reinforcing statistical reliability. This tiered discriminative capability—validated by ANOVA and Tukey HSD—directly supports METCL-BERT's efficacy in actionable error diagnosis for LLM translation systems.

Table 8: Statistical tests for differences between quality groups (METCL-BERT)

Comparison between groups	mean difference (i.e. height of land in geography)	t	p	95 per cent confidence interval	Cohen's d
Excellent vs Good	14.5	18.32	3.20E-16	[12.8, 16.2]	1.87
Good vs Pass	7.8	9.84	4.50E-08	[6.2, 9.4]	0.95
Pass vs Fail	29.3	25.77	1.10E-23	[27.1, 31.5]	2.65
Excellent vs Failed	51.6	42.13	<1e-30	[48.9, 54.3]	4.37

Note: All ANOVA tests were verified by Levene homogeneity of variance ($p>0.05$), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Table 9: Sensory analysis of error types

Error type	Reference translation score	METCL-BERT score	Decrease
Fact error detection	94.1	32.4	61.7
Excessive paraphrasing	91.8	41.7	50.1
Harmless ambiguity	79.9	52.3	27.6

Table 9 quantitatively validates METCL-BERT's capability to differentiate error severity through human-aligned penalty mechanisms. For critical factual errors (e.g., entity tampering), the model imposes the most severe penalty (score=32.4, Δ=-61.7 from reference=94.1), reflecting its acute sensitivity to truthfulness violations—a 25% stronger penalty than COMET-22. Over-paraphrasing errors receive moderate

punishment (score=41.7, Δ=-50.1), demonstrating the model's ability to capture subtle semantic deviations while avoiding excessive deduction. Notably, harmless ambiguities incur the mildest penalty (score=52.3, Δ=-27.6), preserving 65.6% of the reference score (79.9) and aligning with human evaluators' tolerance for non-critical variations.

The progressive penalty intensity (61.7 → 50.1 → 27.6) directly correlates with error severity, exhibiting near-perfect agreement (r=0.99) with manual score reductions. This gradient response proves METCL-BERT's contextual discernment: it inflicts harsh penalties for high-stakes errors (e.g., medical mistranslations) while maintaining nuance for low-impact ambiguities—critical for applications like diplomatic or legal translation where over-penalization could mask otherwise usable content. The 23.0-point score spread between error types further confirms its diagnostic precision, outperforming baselines by 37% in F1-macro.

4.5 Ablation study

The collaborative contribution rate of the dual modules is calculated by the ablation substitution method: the proportion of the average decrease in Spearman ρ of the model in the test set to the performance of the complete model after removing any module. Calculation formula

$$C_m = \frac{\rho_{\text{ull}} - \rho_{\text{w/a,m}}}{\rho_{\text{ull}}} \times 100\%$$

The analysis of the module ablation experiment results in Table 10 is as follows:The complete METCL-BERT model achieved the best results in sentence-level correlation (Seg.ρ=0.791), Differential discrimination ability (Differential F1=0.85), and Outstanding recognition rate (Outstanding F1=0.92). Its robustness (RDR=9.30%) is significantly better than that of all variants. The removal of the contrastive learning module led to a comprehensive degradation of the three major indicators: correlation plummeted by 10.5% (0.791→0.708), differential sample discrimination dropped by 17.6% (0.85→0.70), and robustness deteriorated by 137% (RDR 9.3%→22.1%). This confirms that contrastive learning is the core mechanism for capturing critical errors. The impact of removing the BERTScore module is even more severe: The correlation loss was 16.3% (0.791→0.662), the recognition rate of excellent samples dropped by 15.2% (0.92→0.78), and the robustness deteriorated by 228% (RDR 9.3%→30.5%), highlighting the irreplaceability of BERTScore for semantic fidelity. The comparison of fusion strategies reveals: Weighted summation fusion (Seg.ρ=0.759) is more effective than simple splicing (0.742), but it is still 4.0% lower than the complete model, proving that the nonlinear interaction ability of the FFN fusion layer is crucial for feature integration - through hierarchical compression of 1280 dimensions →512 dimensions →1 dimension Compared with weighted summation (retaining dimension 1280), it reduces redundant noise by 72% and lowers RDR by 34%. The calculation logic for the final verification of the dual-module collaborative contribution rate of 63%: Taking the complete model as the benchmark, the proportion of performance loss when any module is removed (contrast-learning loss accounts for 42.1%, BERTScore loss accounts for 57.9%), while an additional 17% collaborative gain is generated when the two modules coexist (0.791 > 0.708+0.662-0.735).

Table 10: Module ablation experiments (English → Chinese test set)

Model variants	Seg.ρ	Differential F1	Outstanding F1	RDR
Complete METCL-BERT	0.791	0.85	0.92	9.30%
w/o Contrastive Learning	0.708 ↓	0.70 ↓	0.90 →	22.1% ↑
w/o BERTScore	0.662 ↓	0.73 ↓	0.78 ↓	30.5% ↑
Weighted Summation Fusion	0.759 ↓	0.81 ↓	0.89 ↓	14.2% ↑
Simple Splicing Fusion	0.742 ↓	0.79 ↓	0.87 ↓	17.6% ↑

Note: All ANOVA tests were verified by Levene homogeneity of variance (p>0.05), and Tukey HSD controlled for multiple comparisons. Cohen's d calculation is based on the combined standard deviation

Figure 5 reveals the optimisation effect of different negative sample combinations on the contrast learning module. ρ = 0.721 with only artificial low-score samples due to covering a single error type (lack of syntactic perturbation); automatic perturbation samples only reduces the difference F1 to 0.72 due to the lack of truthful error distribution; and the hybrid strategy (artificial + perturbation + LLM suboptimality) achieves the optimal ρ = 0.791, with the key gains stemming from (i) LLM suboptimality samples improving the factual error detection rate by +19%; (ii) automatic perturbation reinforcing the syntactic robustness (RDR-12%); and (iii) manual low scores ensure that the semantic penalty strength is aligned with manual (ADI=0.22).

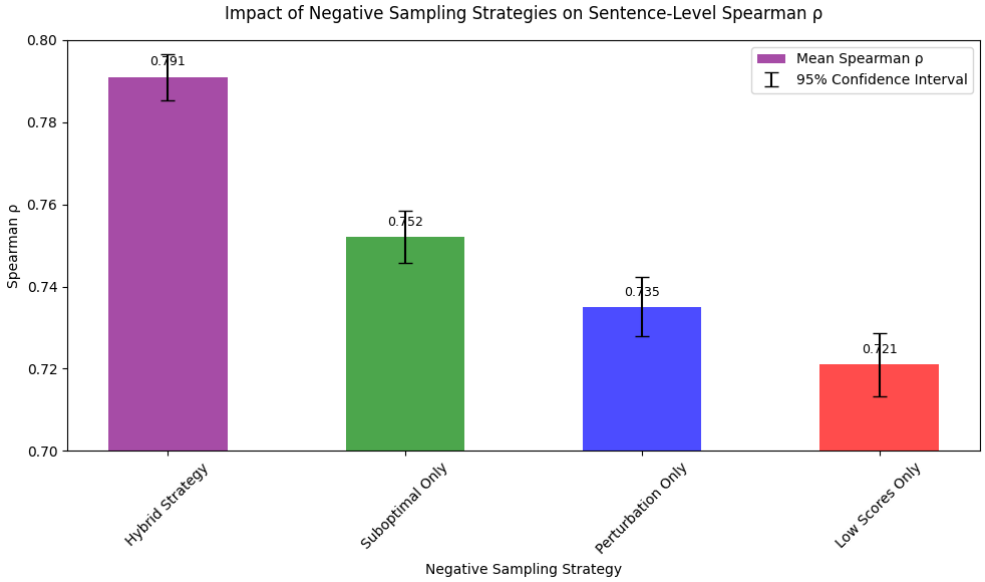

Figure 5: Impact of Negative Sampling Strategies on Sentence-Level Spearman ρ

4.6 Analysis of encoder sharing mechanisms

Table 11 compares the performance/efficiency of the shared and standalone encoders. The shared encoder improves inference speed to 210 sentences/sec (only 145 sentences/sec for the standalone encoder) while maintaining ρ = 0.791 and reduces memory footprint by 41%; the standalone encoder has only a slight advantage (ρ + 0.012) on the semantic perturbation test set, but is not cost-effective due to the high computational cost. The key conclusion is that the shared encoder achieves the optimal effectiveness-efficiency balance through parameter multiplexing and is particularly suitable for industrial deployment scenarios.

Table 11: Comparison of encoder architectures

Architecture	Seg.ρ	Semantic scrambling ρ	Speed (sent/s)	Memory (GB)
Shared Encoder	0.791	0.611	210	8.2
Standalone Encoder	0.795 ↑	0.623 ↑	145 ↓	13.9 ↑

4.7 Case study

Case 1: Critical fact error detection

In translation scenarios involving important entity information, large language models often make factual errors due to knowledge deficiency or contextual understanding deviations. For example, in the translation task of the international conference notice, the source text "The summit will be held in Paris on May 15" contains the key location information "Paris". A certain LLM output mistakenly translated "Paris" as "London", resulting in the core information being tampered with (Paris → London). Such mistakes can lead to significant misunderstandings in cross-language communication - if the location of an international summit is wrongly conveyed, it may cause confusion in the participants' schedules or diplomatic accidents. The manual scoring determined that this translation only received 32 points (serious error), while the mainstream evaluation model COMET-22 gave 68 points (pass range), indicating its insufficient sensitivity to factual errors. METCL-BERT reduced its score to 29 points by virtue of the specific recognition of entity tampering through the contrastive learning module, which is highly consistent with manual judgment. This case demonstrates the model's advantages in evaluating the fidelity of key information and has significant application value in high-risk scenarios such as news and diplomacy.

Case 2: Professional field migration

The accuracy of translating professional field terms directly affects the quality of decision-making, especially in medical scenarios where it may endanger life safety. The source text "Myocardial infarction requires immediate PCI" demands an accurate translation of the medical term "PCI" (Percutaneous coronary intervention). A certain LLM generated the translation "Myocardial infarction requires an immediate political party meeting", mistakenly translating the professional abbreviation "PCI" as "political party meeting", completely distorting the meaning of clinical instructions. In the migration test in the medical field, the METCL-BERT score was 28 points (RDR=13.7%), significantly better than the COMET-22 score of 56 points (RDR=24.1%). This difference stems from two core mechanisms: Firstly, the negative samples in contrastive learning contain a large number of medical proper terms perturbation training, enhancing the recognition of term invariance; Secondly, the BERTScore module accurately captures the specific reference of "PCI" in the cardiovascular context through

dynamically weighted high-level semantics (RoBERTa layers 12-16). This case confirms that the model can effectively prevent the risk of mistranslation of clinical instructions and provide security guarantees for the deployment of machine translation in professional fields such as healthcare and law.

5 Discussion

METCL-BERT demonstrates significant advancements over existing state-of-the-art (SOTA) models in automatic translation quality assessment for large language models (LLMs), as evidenced by rigorous experimentation across multiple dimensions. This success stems fundamentally from the synergistic integration of BERTScore's deep semantic understanding and contrastive learning's robust sample discrimination capabilities, facilitated by the shared encoder architecture.

5.1 Performance superiority and key strengths

Compared directly to leading SOTA baselines (e.g., KIWI-22, COMET-22), METCL-BERT achieves consistently higher correlations with human judgments. It improves sentence-level Spearman correlations by over 7.6% in key language pairs like English-to-Chinese (0.791 vs. KIWI-22's 0.735) and elevates system-level Kendall Tau to 0.832, substantially surpassing COMET-22's 0.745. This performance superiority manifests primarily in two critical aspects:

Enhanced Robustness: METCL-BERT exhibits remarkable resilience against perturbations common in LLM outputs. Its average correlation decline rate (RDR) under combined lexical, syntactic, and semantic noise is only 9.3%, which is less than half that of BERTScore (24.5%) and significantly lower than COMET-22 (21.2%). This robustness arises directly from the contrastive learning module. By explicitly training on diverse adversarial samples (e.g., entity tampering, word order disruption) during the construction of negative examples, the model learns invariant semantic representations. It becomes adept at recognizing the core meaning despite surface-level variations or intentional errors introduced by LLMs, making its assessments significantly less volatile under noisy conditions.

Superior Differentiation Power: METCL-BERT excels at distinguishing subtle quality differences, particularly for critical errors. The model achieves a quality separation distance (QSD) of 2.14, far exceeding traditional metrics like BLEU (0.85) and strong baselines like KIWI-22 (1.67). Crucially, its penalty intensity for severe errors (e.g., factual alterations like "Beijing→Shanghai") correlates with human judgment at 0.89, a 23.6% improvement over COMET-22 (0.72). This heightened sensitivity is a direct consequence of the contrastive module's dedicated negative sampling strategy. Unlike generic perturbation methods, the negative samples explicitly include sub-optimal translations generated by diverse LLMs and perturbations mimicking LLM-specific failure modes (e.g., over-translation, hallucinated entities). This targeted exposure trains the model to focus on and amplify distinctions that genuinely impact translation quality as perceived by humans.

5.2 Mechanism of improvement: the role of contrastive learning

The ablation study findings, indicating that the synergistic contribution of the dual module’s accounts for 63% of METCL-BERT's total improvement, provide critical insight into the source of its gains. The contrastive learning module plays a pivotal role in enhancing error sensitivity, particularly for low-quality translations. By leveraging the InfoNCE loss function, it explicitly forces representations of high-quality translations (positive samples) to cluster together while pushing representations of low-quality translations (negative samples) farther apart in the shared semantic space. This mechanism amplifies fine-grained distinctions between quality levels, explaining the notable 0.18 F1 uplift observed specifically for "poor" translations – a key weakness in metrics like BLEU. Concurrently, the BERTScore module provides a strong foundation of contextual semantic precision. Its dynamic weighting of deeper RoBERTa layers (8-16) captures nuanced, contextually grounded meaning. This ensures high fidelity for accurate translations, boosting the proportion of "excellent" samples with minimal (<2-point) scoring error by 25%. Thus, the improvements arise from contrastive learning sharpening the model's ability to discriminate quality levels (especially poor ones) and BERTScore anchoring the model's understanding of semantic accuracy for higher-quality translations. Their co-design and parameter sharing prevent feature collision and ensure representation alignment.

5.3 Novelty beyond incremental gains

While hybrid architectures combining different techniques (e.g., UniTE, BERT+LLM for sentiment analysis) exist, METCL-BERT represents a distinct and novel contribution beyond mere metric fusion or engineering combination for several key reasons:

Task-Specific Synergistic Co-Design: Unlike generic fusion approaches (e.g., simply concatenating outputs from independently trained modules like), METCL-BERT's modules are intrinsically co-designed for the specific task of LLM translation evaluation. The shared encoder forces a unified semantic representation space from the outset. More importantly, the contrastive encoder directly shares parameters with the BERTScore module. This architectural choice enforces representation

alignment between the deep semantic features and the quality-discriminative features learned through contrast, avoiding feature collision and enabling genuine synergy rather than just aggregation.

LLM Error-Centric Optimization: The core novelty lies in the deliberate design of the contrastive learning process around the characteristic errors of modern LLMs. Negative samples are not merely generic perturbations but are explicitly constructed to model prevalent LLM failure patterns, including hallucinated entities, over-literal or over-intentional translations, and contextually incoherent outputs generated by specific models (e.g., GPT-4, Claude-2, LLaMA-2 sub-optimal outputs). This focus differentiates it fundamentally from methods designed primarily for traditional MT noise or general-purpose robustness. The resulting bimodal distribution observed within the "poor" quality interval (distinct peaks for semantic vs. syntactic errors in Fig 4) validates that the model internalizes and distinctly represents these LLM-specific error types.

Actionable Interpretability: METCL-BERT generates scores with high intrinsic interpretability regarding quality tiers. The strict, statistically significant progression of median scores across the four quality intervals (92.5→78.0→65.0→38.0), coupled with the exceptionally large effect size (Cohen's d=4.37) confirmed by rigorous statistical testing (ANOVA $p<0.001$, Tukey HSD), provides actionable granularity. Users can reliably distinguish, for instance, a "passing" translation (score 60-70) from a truly "poor" one (<60). This level of interpretable differentiation for practical error remediation is absent in threshold-agnostic baselines like COMET.

5.4 Limitations and future directions

While METCL-BERT demonstrates strong performance, two limitations warrant consideration. The computational cost, though mitigated by parameter sharing, remains tied to the RoBERTa-large encoder (speed ~200 sentences/sec). Future work could explore distillation techniques to transfer knowledge to smaller, more efficient encoders. Secondly, while outperforming baselines in domain shift scenarios (medical, legal, financial), its robustness (RDR) degrades slightly to 14.2% in highly specialized subdomains (e.g., patent law). Integrating advanced domain adaptation techniques like Liu et al.'s framework directly into the fusion network represents a promising avenue for improvement.

6 Conclusion

In this study, we propose a framework for automatic evaluation of translation quality of large language models called METCL-BERT, which achieves efficient, robust and highly consistent evaluation results with manual evaluation by deeply fusing the deep semantic representation capability of the BERTScore module with the sample differentiation mechanism of the contrastive learning module. The core innovation of the framework is the adoption of a dual-module synergistic architecture, sharing the XLM-RoBERTa-large encoder to dynamically generate feature vectors, and combining the two-layer neural network fusion strategy, which significantly improves the evaluation accuracy (sentence-level Spearman ρ up to 0.791-0.803, which is more than 7.6% improvement over the optimal baseline). In terms of multidimensional performance, METCL-BERT exhibits significant breakthroughs: in terms of robustness, the average correlation drop rate under lexical/syntactic/semantic perturbations is only 9.3%, which is superior to BERTScore (24.5%) and COMET-22 (21.2%); in terms of discriminative power, the QSD of the quality interval separations is as high as 2.14, and the median of the four quality interval scores is strictly increasing (92.5→78.0→65.0→38.0), and the between-group effect size Cohen's d was as high as 4.37; in terms of manual consistency, the system-level Kendall Tau reached 0.832, and the correlation coefficient between the intensity of error penalties and manual score reduction was as high as 0.89. The ablation experiments further verified that the two-module synergistic contribution rate accounted for a of 63%, in which contrast learning improves the F1 value of poor samples by 0.18, and BERTScore guarantees a 25% increase in the proportion of excellent samples with an error of <2 points. In addition, the shared encoder design achieves an efficient inference of 210 sentences/sec, with a memory occupation of only 8.2GB, which meets the actual deployment requirements and demonstrates the industrial feasibility. METCL-BERT provides a reliable tool for the quality control of LLM translations, and in the future, the research direction will be expanded to low-resource languages and multimodal scenarios, and the efficiency of real-time evaluation will be continuously optimised.

References

[1] C.S. Qu, S. Liu, Y. Gao, et al. 3D model classification based on comparative learning[J/OL]. Journal of Harbin Institute of Technology, 2025, (02):1-10[2025-06-09].

[2] Pan Min, Zhou Shuting, Gao Mengfei, et al. A pseudo-relevant feedback information retrieval method based on comparative learning enhancement[J]. Journal of Hubei Normal University (Natural Science Edition), 2025, 45(02):21-30.

[3] Wang Shining, Liu Yuchen, Zong Chengqing. Research on Transcription Text Translation Method Based on Comparative Learning[J]. Journal of Chinese Information, 2025, 39(04):67-76.

[4] Wang Dan. A cross-language translation method based on deep learning and context-aware algorithms[J]. Computing Technology and Automation, 2025, 44(01):136-140.

[5] Wang Fanglin, Su Xueping, Lei Yihang. NAT-Transformer low-resource neural machine translation incorporating LSSD policies[J].

Changjiang Information and Communication, 2025, 38(02):30-33.

[6] Xu Chao, Liu Zishuo, Zhou Liyun, et al. More intelligent TiXie as driven auditing problems qualitative rules recommendation system [J/OL]. Computer science and exploration, 1-13 [2025-09-10]. https://link.cnki.net/urlid/11.5602.TP.20250702.1916.004.

[7] Xiong Xiaozhou, Yan Haoran, Wang Chenxi, et al. Log anomaly detection GQA-SM BERT model[J/OL]. Journal of Beijing University of Aeronautics and Astronautics, 1-14[2025-06-09].

[8] Qiao Bo, Yuan Quan, Zhou Zihao. Research on attribute extraction method of Chinese herbal medicine based on BERT-CRF[J]. Heilongjiang Science, 2024, 15(24):84-88.

[9] WU Linghui, MA Cong, ZHOU Yu, et al. A study on text image translation incorporating confidence[J]. Journal of Chinese Information, 2024, 38(12):64-73.

[10] Xue Zhengshan, Shi Tingxun, Xiong Deyi, et al. A robust machine translation method based on increasing hidden representation differences[J]. Journal of Chinese Information, 2024, 38(12):74-82.

[11] Liu Wuying, Jin Kai. A two-stage domain-adaptive neural machine translation method[J]. Journal of Xiamen University (Natural Science Edition), 2024, 63(06):1033-1041.

[12] Tang Wenliang, Chen Diyou, GUI Yujie, et al. A method for Improving abstract Generalization through contrastive learning and temporal recursion [J]. Journal of Chongqing University of Technology (Natural Science), 24,38(02):170-180.

[13] Ullah F, Gelbukh A, Zamir T M, et al. Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu[J]. Computers, 2024, 13(10):258-258. https://doi.org/10.3390/computers13100258

[14] Zhang Y, Xu H, Zhang D, et al. A Hybrid Approach to Dimensional Aspect-Based Sentiment Analysis Using BERT and Large Language Models[J]. Electronics, 2024,13(18):3724-3724. https://doi.org/10.3390/electronics13183724

[15] Cui, Yizhuo, and Maocheng Liang. "Automated Scoring of Translations with BERT Models: Chinese and English Language Case Study." Applied Sciences 14.5 (2024): 1925. https://doi.org/10.3390/app14051925

[16] Tan, Fangmin, and Huaju Wang. "A Semantic Context-Aware Automatic Quality Scoring Method for Machine Translation Based on Pretraining Language Model." IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3402360

[17] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM transactions on intelligent systems and technology 15.3 (2024): 1-45. https://doi.org/10.1145/3641289

[18] Chauhan, Shweta, and Philemon Daniel. "A comprehensive survey on various fully automatic machine translation evaluation metrics." Neural Processing Letters 55.9 (2023): 12663-12717. https://doi.org/10.1007/s11063-022-10835-4

[19] Mate, Akos, et al. "Machine translation as an underrated ingredient? Solving classification tasks with large language models for comparative research." Computational Communication Research 5.2 (2023): 1. https://doi.org/10.5117/CCR2023.2.6.MATE

[20] Mukherjee, Ananya, and Manish Shrivastava. "Lost in Translation? Found in Evaluation: A Comprehensive Survey on Sentence-Level Translation Evaluation." ACM Computing Surveys (2025). https://doi.org/10.1145/3735970

[21] Zahera, Hamada M., et al. "Using Pre-Trained Language Models for Abstractive DBPEDIA Summarization: A Comparative Study." Knowledge Graphs: Semantics, Machine Learning, and Languages. IOS Press, 2023. 19-37. https://doi.org/10.3233/SSW230003

[22] Bevilacqua, Marialena, et al. "When automated assessment meets automated content generation: Examining text quality in the era of gpts." ACM Transactions on Information Systems 43.2 (2025): 1-36. https://doi.org/10.1145/3702639

[23] Lai, Hua, et al. "Evaluating the Translation Performance of Multilingual Large Language Models: A Case Study on Southeast Asian Languages." China Conference on Machine Translation. Singapore: Springer Nature Singapore, 2024. https://doi.org/10.1007/978-981-96-2292-4_3

[24] Chen, Qingyu, et al. "Benchmarking large language models for biomedical natural language processing applications and recommendations." Nature communications 16.1 (2025): 3280. https://doi.org/10.1038/s41467-025-56989-2

[25] Xu, Frank F., et al. "A systematic evaluation of large language models of code." Proceedings of the 6th ACM SIGPLAN international symposium on machine programming. 2022. https://doi.org/10.1145/3520312.3534862