When Correct Translation Is Not Enough: Learning Human-Like Quality Judgments for Data Selection in Low-Resource Machine Translation

Anonymous ACL submission

Abstract

Low-resource machine translation (MT) often relies on synthetic parallel data generated by large language models and conventional MT systems. However, existing data selection methods mainly focus on semantic consistency or distributional signals, often overlooking stylistic naturalness and idiomatic expression. As a result, translations that are adequate and fluent may still sound rigid, literal, or culturally unnatural.We model translation quality for low-resource MT from three human-inspired dimensions: adequacy, fluency, and elegance. We construct an expert-annotated multi-dimensional scoring dataset and an elegance-oriented preference dataset, and propose *TransJudge*, a lightweight reward model trained with supervised fine-tuning and preference optimization for scalable translation quality evaluation and data selection.Experiments on Indonesian–Chinese MT show that TransJudge achieves stronger agreement with human judgments than automatic metrics and competitive agreement with LLM-as-judge baselines, while reducing evaluation latency by up to 50× and eliminating per-query API costs. For downstream data selection, TransJudge improves MT performance by 1.84 BLEU and 2.31 COMET on average over baseline selection methods.

1 Introduction

> *The three difficulties in translation are: adequate, fluent, and elegant.*
>
> – Yan, 1898

Instruction tuning has been shown to significantly improve the translation capabilities of large language models (LLMs) (Grattafiori et al., 2024; Bai et al., 2023), especially for high-resource language , but it relies on the availability of sufficient high-quality parallel corpora. In low-resource machine translation (MT), the scarcity of human-annotated parallel data constrains model learning. To alleviate this limitation, synthetic translations generated by LLMs or conventional MT systems are widely used to augment training corpora (Kreutzer et al., 2022; He et al., 2023). Nevertheless, synthetic data are often noisy and heterogeneous, containing semantic inconsistencies, alignment errors, or unnatural target-language expressions. Training on such low-quality samples can directly impair downstream translation performance (Khayrallah and Koehn, 2018; Koehn and Knowles, 2017). Consequently, **data selection** has emerged as a crucial strategy to identify high-quality samples and ensure effective model training in low-resource MT .

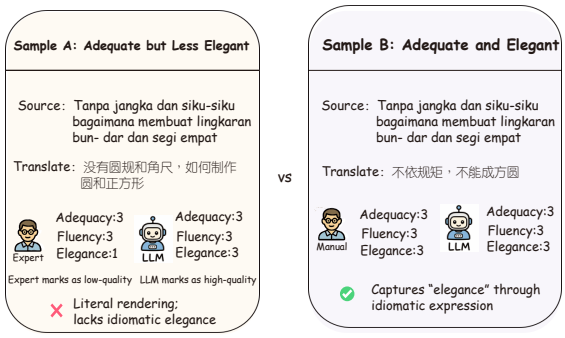

Figure 1: LLM may overestimate elegance when translations are adequate and fluent

Existing translation data selection methods usually rely on signals such as perplexity, semantic similarity, model confidence, or data diversity (Pan et al., 2024; Thrush et al., 2025). These signals are useful for estimating whether a sentence pair is semantically aligned or distributionally useful, but they provide limited evidence about whether the target sentence is natural, idiomatic, and stylistically appropriate. As a result, existing methods may retain translations that are meaning-preserving and readable, yet still too literal or unnatural to serve as high-quality training examples. This limitation raises a central question: *what kind of translation*

data should be considered truly high-quality for low-resource MT?

Motivated by the classical translation principle of "faithfulness, expressiveness, and elegance" (Fu, 1898), we define translation quality through three complementary dimensions: **adequacy**, **fluency**, and **elegance**. Adequacy captures faithfulness to the source meaning, fluency captures grammaticality and readability, and elegance captures idiomatic, natural, and stylistically appropriate expression in the target language. Together, they provide a broader view of translation quality, covering semantic correctness, linguistic well-formedness, and human-like expression beyond literal adequacy.

However, using these dimensions for data selection is non-trivial, especially for stylistic elegance. Unlike adequacy and fluency, which can be partially assessed through semantic preservation and grammatical well-formedness, elegance depends on subtler properties such as idiomatic expression, cultural naturalness, and human-like phrasing. These properties are difficult to quantify, particularly in low-resource settings where high-quality human translations and expert judgments are scarce. A straightforward solution is to use LLMs as judges to score translations along the three dimensions. Yet, as illustrated in Fig. 1, LLM-as-judge systems may overestimate translations that are adequate and fluent but still rigid, literal, or stylistically unnatural, since their judgments are often dominated by semantic correctness and surface fluency. Moreover, repeatedly querying proprietary LLM APIs for large-scale synthetic corpus filtering incurs substantial computational and financial costs.Beyond evaluating each dimension separately, data selection further requires integrating adequacy, fluency, and elegance into a unified utility score that reflects the expected training value of each synthetic sentence pair. These limitations motivate two key challenges:

(C1) How can we obtain reliable and learnable supervision for adequacy, fluency, and especially stylistic elegance in low-resource MT data selection? **(C2)** How can we distill multi-dimensional translation quality signals into a lightweight and scalable reward model that estimates the utility of synthetic sentence pairs without repeated reliance on expensive LLM inference?

To address **C1**, we construct an expert-annotated dataset for low-resource translation, containing 5,000 sentence pairs with fine-grained scores on adequacy, fluency, and elegance. To better capture the subjective nature of elegance, we further build an elegance-oriented preference dataset with 3,077 preference pairs and use DPO-style preference learning to distinguish idiomatic, human-like translations from rigid literal ones.

To address **C2**, we propose *TransJudge*, a lightweight reward model trained jointly with multi-dimensional scoring supervision and elegance-oriented preference supervision. TransJudge maps adequacy, fluency, and elegance into a unified utility score for each synthetic sentence pair, enabling scalable filtering of high-value training data for downstream low-resource MT without repeated reliance on expensive LLM inference.

To summarize, we make the following contributions:

(1) We model translation quality in low-resource MT from three dimensions: adequacy, fluency, and elegance, and use them as supervision signals for translation data selection.

(2) We construct an expert-annotated multi-dimensional evaluation dataset together with an elegance-oriented preference dataset, enabling both fine-grained quality assessment and comparative supervision for stylistic naturalness.

(3) We propose *TransJudge*, a lightweight reward model trained with supervised fine-tuning and DPO-style preference optimization, providing a scalable alternative to costly LLM-as-judge evaluation for large-scale synthetic data filtering.

(4) Extensive experiments demonstrate that TransJudge achieves stronger agreement with human judgments and consistently improves downstream MT performance over existing data selection baselines. TransJudge further enables substantially more efficient large-scale filtering, reducing evaluation latency by up to 50× and eliminating API cost compared with GPT-4-based evaluation.

2 Method

2.1 Framework Overview

The overall framework of *TransJudge* is illustrated in Fig. 1. Our goal is to transform abstract translation quality principles into learnable reward signals for scalable low-resource MT data selection. To achieve this, the framework consists of two stages: (a) multi-dimensional reward learning and (b) adaptive reward fusion for data selection.

In Stage A, we first construct expert-annotated supervision data covering three dimensions of translation quality: adequacy, fluency, and elegance

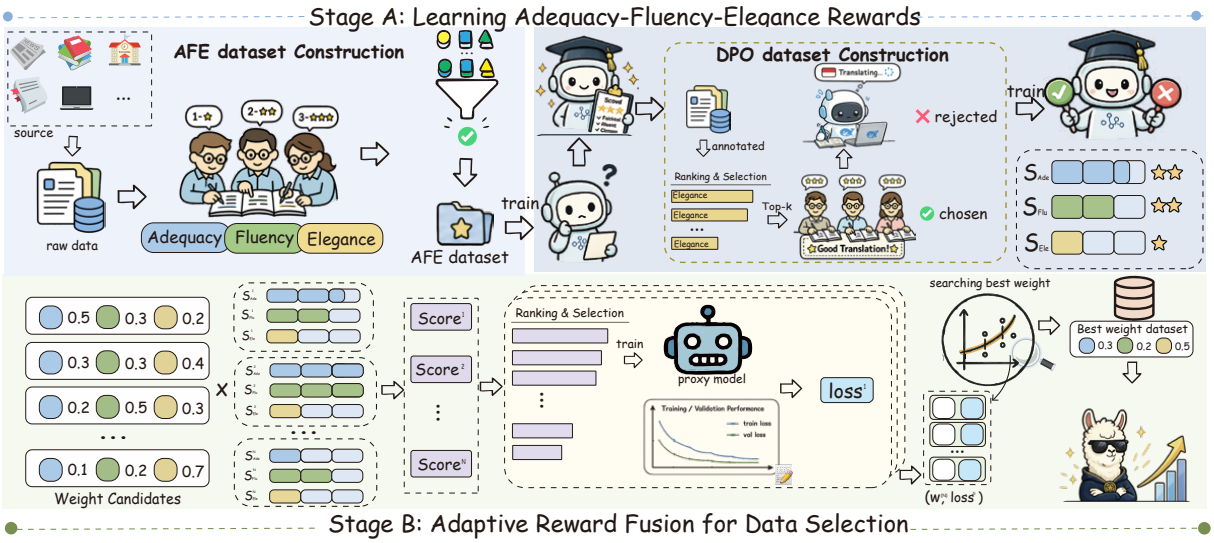

Figure 2: Overview of the TransJudge framework. Stage A performs AFE-based reward modeling with multi-dimensional supervision and preference learning, while Stage B conducts proxy-guided reward fusion and data selection for downstream MT training.

(AFE). Based on this annotation scheme, we build a multi-dimensional scoring dataset and a preference dataset targeting stylistic quality. We then train a lightweight reward model on the constructed supervision data, enabling scalable evaluation across adequacy, fluency, and elegance.

In Stage B, the trained reward model is used to score large-scale synthetic parallel data from multiple quality dimensions. Instead of manually assigning fixed weights to different reward dimensions, we introduce an adaptive reward fusion strategy that searches for effective weight combinations through proxy-model training. The resulting fused reward is used to rank and filter synthetic parallel data for downstream MT training.

Through this design, TransJudge provides a scalable and low-cost alternative to expensive LLM-based evaluation while enabling more effective selection of high-value synthetic translation data.

2.2 Dataset Construction

2.2.1 AFE Annotation Scheme

Existing automatic evaluation metrics for machine translation mainly focus on semantic correctness or sentence-level similarity, but often fail to capture whether a translation is natural, idiomatic, and culturally appropriate. In low-resource MT, such distinctions are especially important, since synthetic translations may be semantically acceptable while still exhibiting rigid or overly literal expressions.

To better characterize translation quality, we adopt a three-dimensional annotation framework inspired by the classical translation principle of adequacy, fluency, and elegance (AFE). This framework operationalizes translation quality into learnable supervision signals for reward modeling and data selection.

Adequacy measures whether the translation faithfully preserves the original meaning, including semantic consistency, omissions, and mistranslations. **Fluency** evaluates the readability and grammatical well-formedness of the target sentence, focusing on coherence, word order, and overall comprehensibility. **Elegance** further measures stylistic naturalness and idiomatic expression, emphasizing whether the translation conforms to native Chinese usage through appropriate wording, concise phrasing, culturally suitable expression, and natural rendering of idioms or figurative language. A translation may be adequate and fluent while still sounding overly literal or mechanically translated.

To improve annotation consistency, each dimension is scored independently using a three-level scale: **1 (Poor)**, **2 (Fair)**, and **3 (Good)**. Compared with finer-grained scoring schemes, the three-level setting reduces subjective ambiguity and improves inter-annotator consistency in low-resource translation evaluation.

All samples are independently annotated by three annotators with Indonesian–Chinese bilingual backgrounds. Before formal annotation, annotators complete a pilot annotation stage to calibrate scoring standards. During annotation, annotators are instructed to evaluate the three dimensions sepa-

rately rather than assigning template-like scores. To further ensure annotation quality, we introduce an expert review mechanism for high-variance samples. Specifically, samples with large scoring disagreement are re-examined by senior reviewers, and the final labels are determined through expert adjudication. Detailed annotation guidelines and scoring examples are provided in Appendix A.

2.2.2 Multi-Dimensional Supervision Data

To train the reward model with diverse translation quality signals, we construct a multi-dimensional supervision dataset covering a broad range of translation phenomena in Indonesian–Chinese MT. Rather than relying on a single corpus or domain, we collect translation pairs from diverse bilingual resources, including dictionary-derived bilingual entries, web-crawled parallel texts, news-domain corpora, and manually collected translation examples. These resources contain substantial variation in sentence structure, expression style, and translation difficulty.

During dataset construction, we intentionally preserve diversity in linguistic phenomena and stylistic patterns instead of focusing only on fluent or easily translatable examples. In particular, the dataset includes idioms, colloquial expressions, culturally grounded phrases, literal translations, and formal news-style sentences. This diversity improves coverage of translation phenomena in reward learning.

All samples are annotated according to the AFE annotation scheme described in Section 2.2.1. The final scoring dataset contains 5,000 Indonesian–Chinese translation pairs annotated with independent adequacy, fluency, and elegance scores. Additional dataset statistics and annotation examples are provided in Appendix A.

2.2.3 Preference Pair Construction

While the multi-dimensional scoring dataset provides pointwise supervision for translation quality assessment, stylistic naturalness and idiomatic expression are often difficult to model through absolute scores alone. In many cases, two translations may both preserve source semantics and remain grammatically understandable, while differing substantially in stylistic quality and naturalness. To better capture such subtle distinctions, we further construct a preference-based supervision dataset for pairwise reward learning.

We first identify stylistically strong translations from the annotated scoring dataset based on their AFE annotations. These translations are primarily derived from human-refined or manually collected translation examples and exhibit natural, idiomatic, and culturally appropriate Chinese expression. We then use the corresponding Indonesian source sentences to generate alternative candidate translations using a large language model following a standard MT prompting setup. Unlike manually constructed negative samples, these LLM-generated translations are usually semantically acceptable and generally understandable, but often contain more literal or stylistically rigid expressions.

To improve preference quality, we manually filter generated candidates and retain only samples where the main distinction lies in stylistic naturalness rather than severe semantic errors or grammatical failures. As a result, the chosen and rejected translations typically remain comparable in adequacy and basic fluency, while differing mainly in idiomaticity and expression quality.Fig 3 presents a representative example from the preference dataset.

The final preference dataset contains 3,077 translation pairs. This preference supervision complements the pointwise scoring dataset by encouraging the reward model to distinguish natural human-like translations from semantically correct but overly literal alternatives. Additional preference examples and generation prompts are provided in Appendix B.

```
instruction: "Translate the following Indonesian sentence into Chinese.
input: "Tempat tinggal bangsawan tak gampang dimasuki orang awam."
chosen: “侯门深似海”
rejected: “贵族的居住地方不容易被普通的人进入。”
```

Figure 3: Representative preference pair used for stylistic preference learning.

2.3 Adequacy-Fluency-Elegance Reward Modeling

In Stage A, we instantiate TransJudge with Qwen2.5-7B-Instruct (Ahmed et al., 2025) and fine-tune it to perform AFE-based translation quality assessment. For each instance, the input consists of a source sentence x and a translation y, and the output contains three scores on a three-point scale, corresponding to adequacy, fluency, and elegance.

We first conduct supervised fine-tuning (SFT) on the 5,000-instance AFE scoring dataset. The training objective is to teach the model to predict adequacy, fluency, and elegance scores for each translation pair. This step enables TransJudge to distinguish semantic adequacy, target-language readability, and stylistic quality as separate reward signals. The resulting SFT model serves as the initialization for the subsequent preference optimization stage.

2.4 Elegance-Oriented Preference Optimization

Since elegance is inherently more subjective and preference-oriented than adequacy and fluency, it is difficult to fully capture through scalar score supervision alone. In many cases, multiple translations may preserve the same source meaning while differing subtly in idiomaticity, stylistic naturalness, and cultural appropriateness. To better model such relative stylistic distinctions, we further refine TransJudge with DPO (Rafailov et al., 2023) on the 3,077-instance elegance preference dataset.

Each preference example is represented as (x, y^c, y^r), where y^c denotes the *chosen* translation and y^r denotes the *rejected* translation. The chosen translation is manually written or refined to be more idiomatic and natural, while the rejected translation is generated by a language model and filtered to remain understandable but less elegant.

The preference objective encourages the model to prefer the chosen translation over the rejected one under the same source sentence, denoted as $y^c \succ_x y^r$. This stage further calibrates TransJudge toward elegance-oriented distinctions beyond pointwise quality scoring.

After SFT and DPO, TransJudge is used as an AFE scorer. At inference time, each candidate pair is formatted with the AFE evaluation prompt, and the model outputs a structured response containing adequacy, fluency, and elegance scores. The full prompt template is provided in Appendix C.

2.5 Proxy-Guided Reward Fusion

After Stage A, TransJudge assigns each candidate translation pair (x, y) a three-dimensional AFE score vector:

$$\mathbf{s}(x, y) = \{s_{\text{Ade}}(x, y), s_{\text{Flu}}(x, y), s_{\text{Ele}}(x, y)\}, \quad (1)$$

where the three dimensions correspond to adequacy, fluency, and elegance. Stage B learns how to fuse these dimension-level scores into a unified reward score for data selection. Instead of manually assigning fixed weights, we adopt a proxy-guided try-and-validate strategy that estimates the contribution of each dimension according to downstream MT performance.

Inspired by MetaRater (Zhuang et al., 2025), we sample multiple candidate weight vectors:

$$\mathbf{w}^{(i)} = \{w_1^{(i)}, w_2^{(i)}, w_3^{(i)}\}, \quad i = 1, 2, \ldots, N. \quad (2)$$

For each weight configuration, the aggregated reward score is computed as:

$$S^{(i)}(x, y) = \sum_{j=1}^{3} w_j^{(i)} s_j(x, y). \quad (3)$$

Based on $S^{(i)}$, we rank the candidate data and select the top-k subset:

$$D_{\text{sel}}^{(i)} = \text{Top-}k\left(S^{(i)}\right). \quad (4)$$

A lightweight proxy MT model $M^{(i)}$ is then trained on $D_{\text{sel}}^{(i)}$. Its validation loss is computed as:

$$L^{(i)} = \mathcal{C}\big(M^{(i)}, D_{\text{val}}\big), \quad (5)$$

where $\mathcal{C}(\cdot)$ denotes the average validation loss on the held-out set. This produces a set of weight–performance pairs:

$$\{(\mathbf{w}^{(i)}, L^{(i)})\}_{i=1}^{N}. \quad (6)$$

We fit a regression function to approximate the relationship between reward weights and downstream validation loss:

$$L \approx f(\mathbf{w}; \theta), \quad (7)$$

where $f(\cdot)$ is the regression model parameterized by θ. The optimal weight vector is obtained by:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w}; \theta). \quad (8)$$

This proxy-guided strategy adaptively calibrates the contribution of adequacy, fluency, and elegance according to downstream MT performance, avoiding manually specified reward composition.

3 Experimental Setup

3.1 Datasets

For downstream data selection, we construct a large-scale Indonesian–Chinese candidate pool from heterogeneous web sources, including naturally occurring bilingual texts and automatically constructed translation pairs. After filtering, the full corpus contains approximately 700K sentence pairs, from which we randomly sample 100K instances for data selection experiments. All methods rank and select training data from the same candidate pool for downstream MT fine-tuning.

3.2 Baselines

We compare TransJudge with several representative data selection baselines, including Full Data (Xia et al., 2024), Random Selection (Xia et al., 2024), PPL (Li et al., 2024a), IFD (Li et al., 2024b), SelectIT (Liu et al., 2024), Token Length (TL) (Xia et al., 2024), and ZIP (Yin et al., 2024).

3.3 Benchmarks and Metrics

Benchmarks and Metrics. To ensure fair comparison and reproducibility, we evaluate all methods on the FLORES-200 benchmark (NLLB Team, 2024), which provides high-quality multilingual evaluation data for low-resource machine translation. Following prior work, we adopt the Devtest splits for the Indonesian–Chinese translation direction while excluding the English side.

We report three widely used machine translation evaluation metrics: (1) BLEU (Papineni et al., 2002), which measures n-gram overlap between generated translations and references; (2) ROUGE-L (Lin, 2004), which evaluates longest common subsequence similarity and reflects surface-level fluency and coverage; and (3) COMET (Rei et al., 2020), a learned evaluation metric that better correlates with human judgment by jointly modeling semantic adequacy and translation quality.

3.4 Implementation Details

We evaluate downstream MT fine-tuning on Qwen2.5-7B (Ahmed et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024). For both models, we use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. The learning rate is set to 1×10^{-5}, and each model is trained for 3 epochs. We set the per-device training batch size to 1 and use gradient accumulation with 8 steps. All experiments are conducted on NVIDIA A30 GPUs.

4 Experimental Results

4.1 Main Results

Experimental results are reported in Table 1. For fair comparison, all data selection methods retain 5% of the candidate pool, while *Full* denotes training on the entire unfiltered dataset. The main findings are summarized as follows.

(1) High-quality selected data can outperform full-data training. Using only 5% of the candidate data, TransJudge consistently surpasses the *Full* baseline on both backbone models. On LLaMA-3.1-8B, BLEU improves from 24.19 to 26.07 and COMET from 77.31 to 81.98. On Qwen2.5-7B, BLEU increases from 25.80 to 28.86, while COMET improves from 76.22 to 80.67. These results indicate that carefully selected high-quality samples are more beneficial than simply increasing training data volume.

(2) TransJudge provides a more reliable filtering signal than existing selection methods. TransJudge achieves the best performance across all metrics and backbone models. Compared with the strongest baseline, SelectIT, it improves BLEU/-COMET by 5.12/3.57 on LLaMA-3.1-8B and by 3.37/2.03 on Qwen2.5-7B. In contrast, methods based on length, perplexity, compression, or instruction-following difficulty often perform inconsistently and sometimes even fall below random selection. This suggests that surface-level or likelihood-based signals are insufficient for measuring translation quality, whereas TransJudge directly captures faithfulness, fluency, and elegance.

(3) TransJudge generalizes across different backbone models. TransJudge consistently improves translation quality on both LLaMA-3.1-8B and Qwen2.5-7B despite their different architectures, demonstrating that the proposed evaluator learns transferable translation quality signals rather than overfitting to a specific model family.

4.2 Effect of Selection Ratio

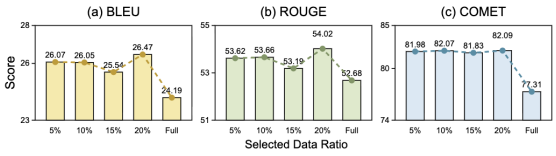

Figure 4: Effect of selected data percentage on LLaMA-3.1-8B under different selection ratios.

We further conduct an exploratory sensitivity analysis on the selected data ratio by varying the retained data from 5% to 20% and comparing it with full-data training. As shown in Figure 4, performance consistently improves as the selected data ratio increases, achieving the best BLEU, ROUGE-L, and COMET scores at 20%. In contrast, training on the full dataset leads to a clear performance drop across all metrics, suggesting that data quality is more important than data quantity in low-resource MT. These results further demonstrate the effectiveness of strategic data selection in our framework, while a more comprehensive study of selection ratios and dataset scales is left for future work.

Backbone	Metric	Full	Random	Length	PPL	ZIP	IFD	SelectIT	TransJudge
LLaMA-3.1-8B	BLEU	24.19	23.62	22.45	21.65	22.72	19.41	20.95	**26.07**
	ROUGE-L	52.68	51.15	50.27	50.07	50.73	47.72	48.55	**53.62**
	COMET	77.31	79.15	76.85	77.14	76.96	74.67	78.41	**81.98**
Qwen2.5-7B	BLEU	25.80	27.07	25.28	24.41	24.74	24.41	25.49	**28.86**
	ROUGE-L	54.38	54.84	53.53	53.20	52.95	53.20	53.06	**56.32**
	COMET	76.22	78.88	76.19	76.42	77.16	76.42	78.64	**80.67**

Table 1: Main results on Indonesian–Chinese translation. Best in **bold**; second-best underlined.

4.3 Dimension Ablation

Variant	BLEU	R-L	COMET	Avg
Only Adequacy	24.95	52.56	81.56	53.02
Only Fluency	22.83	50.55	80.68	51.35
Only Elegance	23.40	51.04	80.88	51.77
w/o Adequacy	24.12	51.68	81.28	52.36
w/o Fluency	25.74	53.14	81.98	53.62
w/o Elegance	24.92	52.27	81.52	52.90
TransJudge (Ours)	**26.07**	**53.62**	**81.98**	**53.89**

Table 2: Ablation study of AFE reward dimensions on LLaMA-3.1-8B. "Only" uses one dimension, and "w/o" removes one. R-L: ROUGE-L.

To examine the contribution of each judging dimension, we compare the full TransJudge with variants that either use a single dimension or remove one dimension from the scoring function. The results are shown in Table 2. Among the single-dimension variants, faithfulness provides the strongest signal because it directly reflects semantic adequacy. However, relying only on faithfulness may overlook readability and stylistic naturalness, while fluency or elegance alone cannot fully ensure meaning preservation. The full TransJudge achieves the best overall balance by combining complementary dimensions, showing that robust translation data selection requires more than a single quality criterion.

4.4 Agreement with Human Judgments

To evaluate whether TransJudge aligns with human assessments of translation quality, we measure its correlation with human judgments across faithfulness, fluency, and elegance. Table 3 reports Pearson's r (Sedgwick, 2012), Spearman's ρ (de Winter et al., 2016), and Kendall's τ (McLeod, 2005) correlations between automatic evaluators and aggregated human ratings.

As shown in Table 3, TransJudge achieves the strongest agreement with human judgments on fluency and elegance across all three correlation metrics, demonstrating its ability to capture readability, stylistic naturalness, and idiomatic expression. For faithfulness, Qwen obtains slightly higher rank-based correlations, while TransJudge remains competitive and achieves the highest Pearson correlation jointly with Qwen. These results suggest that the proposed multi-dimensional supervision and DPO-based preference optimization help TransJudge better align with human preferences, particularly for distinguishing natural translations from rigid literal renderings. Inter-annotator agreement statistics are further reported in Appendix E.

Metric	Pearson's r ↑	Spearman's ρ ↑	Kendall's τ ↑
Faithfulness			
TransJudge	**0.75**	0.71	0.67
DeepSeek	0.62	0.62	0.56
Qwen	**0.75**	**0.73**	**0.69**
GPT-4	0.62	0.61	0.56
Fluency			
TransJudge	**0.83**	**0.81**	**0.77**
DeepSeek	0.81	0.71	0.69
Qwen	0.65	0.58	0.55
GPT-4	0.76	0.68	0.65
Elegance			
TransJudge	**0.71**	**0.69**	**0.64**
DeepSeek	0.67	0.63	0.60
Qwen	0.58	0.58	0.54
GPT-4	0.63	0.59	0.55

Table 3: Correlation between automatic evaluators and human judgments across translation dimensions. See Appendix E.3 for visualization.

4.5 Efficiency Analysis

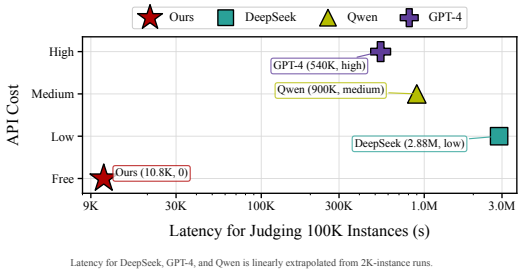

Figure 5: Efficiency comparison of judge models for evaluating 100K translation pairs.

Efficiency is critical for large-scale synthetic data filtering. Although proprietary LLM-as-Judge systems achieve strong agreement with human judgments, they require repeated API-based inference, resulting in substantial cost and latency. As shown in Figure 5, TransJudge evaluates 100K translation pairs in approximately 3 hours using local batch inference without API cost, while maintaining agreement with human judgments comparable to strong LLM evaluators, demonstrating its practicality and scalability for low-resource MT data selection.

To further evaluate generalization beyond Indonesian–Chinese translation, we conduct experiments on Hindi–Chinese MT using a crawled corpus containing 48,016 translation pairs. Following the same setting as the Indonesian–Chinese experiments, we use FLORES-200 as the test benchmark with a 5% selection budget. As shown in Table 4, TransJudge consistently outperforms random selection and existing baselines across BLEU, ROUGE-L, and COMET, improving BLEU from 24.62 to 26.47 and COMET from 78.76 to 81.33 over random selection. These results demonstrate that the proposed AFE-based reward modeling strategy generalizes effectively across different low-resource language pairs.

4.6 Generalization to Other Language Pairs

Method	BLEU	R-L	COMET
Random	24.62	52.00	78.76
SelectIT	24.49	51.96	79.01
TL	24.05	50.71	78.84
TransJudge	**26.47**	**53.95**	**81.33**

Table 4: Generalization results on Hindi–Chinese translation with LLaMA-3.1-8B under a 5% selection budget.

To evaluate the generalization ability of TransJudge beyond Indonesian–Chinese translation, we further conduct experiments on Hindi–Chinese MT using the same 5% data selection setting.

As shown in Table 4, TransJudge consistently outperforms random selection and existing baselines across all evaluation metrics. In particular, it improves BLEU from 24.62 to 26.47 and COMET from 78.76 to 81.33 over random selection.

These results suggest that the proposed AFE-based reward modeling strategy generalizes beyond a single language pair and remains effective for low-resource MT data selection across different linguistic settings.

5 Related Work

Data Selection. Existing data selection methods can be broadly divided into task-agnostic and task-specific approaches. Task-agnostic methods rely on general signals such as perplexity (Marion et al., 2023), gradients (Pan et al., 2024), or data learnability (Zhou et al., 2023), while more recent work explores iterative filtering strategies for multi-perspective quality estimation (Lin et al., 2025). Although effective for removing noisy or redundant data, these methods often fail to capture translation-specific quality characteristics. Task-specific approaches instead tailor selection strategies to downstream applications such as medical (Ding et al., 2025) or legal domains (Li et al., 2025) using domain-aware signals and expert knowledge. However, translation-specific data selection methods for low-resource MT remain limited.

Translation Quality Evaluation. Reliable evaluation is essential for effective translation data selection. Traditional metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) mainly rely on lexical overlap, while neural metrics such as COMET (Rei et al., 2020) model semantic similarity through learned representations. More recently, LLM-as-Judge evaluation (Gu et al., 2024; Umar et al., 2026) has shown strong ability to capture human-like linguistic preferences. Nevertheless, existing metrics remain limited for low-resource MT (Chen et al., 2025). Lexical and neural metrics often overlook stylistic naturalness and idiomatic expression, while LLM-based evaluators suffer from instability and bias under low-resource settings and incur substantial inference cost for large-scale filtering.

6 Conclusion

We propose TransJudge, a lightweight reward model for low-resource MT data selection. TransJudge models translation quality from faithfulness, fluency, and elegance through multi-dimensional supervision and DPO-based preference optimization. Experimental results show that TransJudge achieves strong agreement with human judgments and consistently improves downstream translation performance using only a small subset of candidate data. In addition, TransJudge supports efficient local batch inference, providing a scalable alternative to expensive LLM-as-Judge systems for large-scale synthetic data filtering.

7 Limitations

Although TransJudge provides an effective and scalable framework for low-resource MT data selection, several limitations remain. Our experiments currently cover only a small number of language pairs, while stylistic elegance and idiomatic expression are often language- and culture-dependent, requiring broader evaluation across more diverse low-resource settings. In addition, elegance itself is inherently subjective and may vary across annotators, domains, and target audiences. Although preference optimization helps model relative stylistic preferences, fully capturing human judgments of naturalness and cultural appropriateness remains challenging. Finally, the current framework mainly focuses on sentence-level translation quality, leaving document-level coherence, terminology consistency, and discourse-level style preservation for future work.

Acknowledgments

AI assistants were used for language polishing and improving writing clarity. All technical decisions, experiments, and final verification were conducted by the authors.

References

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Md. Naseef Ur Rahman Chowdhury, and Ahshanul Haque. 2025. Qwen2.5: A comprehensive review of the leading resource-efficient LLM with potential to surpass all competitors.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024.

Joost C. F. de Winter, Samuel D. Gosling, and Jeff Potter. 2016. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273–290.

Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Weibin Liao, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. 3DS: Medical domain adaptation of LLMs via decomposed difficulty-based data selection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19473–19495.

Yan Fu. 1898. *Evolution and Ethics*. Classical work, no publisher information available.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Zexue He, Graeme Blackwood, Rameswar Panda, Julian McAuley, and Rogerio Feris. 2023. Synthetic pre-training tasks for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8080–8098, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, and 2 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024a. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81.

Xiaotian Lin, Yanlin Qi, Yizhang Zhu, Themis Palpanas, Chengliang Chai, Nan Tang, and Yuyu Luo. 2025. LEAD: Iterative data selection for efficient LLM instruction tuning. *Proceedings of the VLDB Endowment*, 19(3):426–439.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. SelectIT: Selective instruction tuning for large language models via uncertainty-aware self-reflection. *arXiv preprint arXiv:2402.16705*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*.

A. Ian McLeod. 2005. Kendall rank correlation and Mann-Kendall trend test. *R Package Kendall*, 602:1–10.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based diverse and high-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15395–15406.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 2685–2702.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *BMJ*, 345.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. 2025. Improving pretraining data using perplexity correlations. In *International Conference on Learning Representations*, pages 17475–17509.

Umar Baba Umar, Sulaimon Adebayo Bashir, and Abdulmalik Danlami Mohammed. 2026. Anchoring the judge: Curriculum-based adaptation and reference-anchored MQM for LLM-based machine translation of an unseen low-resource language: A case of Nupe. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 200–211.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. 2024. Rethinking data selection at scale: Random selection is almost all you need. *arXiv preprint arXiv:2410.09335*.

Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. 2024. Entropy law: The story behind data compression and LLM performance. *arXiv preprint arXiv:2407.06645*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Jiantao Qiu, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10856–10896.

A Annotation Guidelines

A.1 Annotation Procedure

All translations were independently annotated by three bilingual annotators with Indonesian–Chinese translation backgrounds. Annotators were instructed to score each translation separately on adequacy, fluency, and elegance according to the AFE rubric. To reduce annotation bias, all samples were randomly shuffled and independently evaluated without access to other annotators' scores. For samples with large disagreement, annotations were further reviewed and discussed before finalizing the scores. Final scores were obtained by averaging the annotations across evaluators.

A.2 AFE Scoring Rubric

To operationalize translation quality for low-resource MT evaluation, we annotate translations from three dimensions: **Adequacy**, **Fluency**, and **Elegance** (AFE). Existing automatic evaluation metrics mainly focus on semantic correctness or sentence-level similarity, but often fail to capture whether a translation is natural, idiomatic, and culturally appropriate. In low-resource MT, such distinctions are particularly important because synthetic translations may preserve source meaning while still exhibiting rigid or mechanically translated expressions.

We adopt a three-level scoring scheme instead of finer-grained ratings to reduce subjective ambiguity and improve inter-annotator consistency in stylistic evaluation. Each dimension is independently scored as **1 (Poor)**, **2 (Fair)**, and **3 (Good)**.

Table 5 summarizes the scoring criteria used in our annotation process.

A.3 Annotation Examples

Table 6 presents representative examples from our annotation dataset. These examples illustrate that adequacy, fluency, and elegance capture different aspects of translation quality. In particular, some translations may preserve source semantics and remain grammatically understandable while still lacking idiomatic or stylistically natural Chinese expression.

Dimension	Scoring Criteria
Adequacy	**1**: Severe mistranslation, omission, semantic reversal, or factual inconsistency. **2**: Main meaning is preserved with minor semantic deviations or omissions. **3**: Translation accurately preserves the source meaning without major semantic errors.
Fluency	**1**: Translation is grammatically broken, confusing, or difficult to understand. **2**: Translation is understandable but contains awkward phrasing or translationese. **3**: Translation is natural, coherent, and grammatically fluent in Chinese.
Elegance	**1**: Translation is overly literal, mechanically phrased, or unnatural in Chinese expression. **2**: Translation contains partial localization or stylistic adaptation, but remains relatively ordinary. **3**: Translation is idiomatic, stylistically natural, and culturally appropriate for native Chinese readers.

Table 5: AFE annotation rubric used in our dataset construction.

AFE Annotation Examples

Source Sentence	Translation	AFE	Explanation
Mulut manusia tiada yang menjaga.	人的嘴巴没有看守的。	3/2/1	The translation preserves the core meaning but follows an overly literal structure that does not conform to natural Chinese expression.
Mulut manusia tiada yang menjaga.	嘴上没把门儿的。	3/3/3	The translation accurately conveys the original meaning while adopting an idiomatic and culturally natural Chinese expression.
Apakah itu diakses, Anda bisa mengklik dan mendapatkan aplikasi ini luar biasa.	它访问了，您可以单击并获得此非凡的应用程序。	2/1/1	The translation contains broken grammar and unnatural wording, resulting in poor readability and stylistic quality.
Hurufnya menggunakan huruf sirilik.	他的信用完了终点。	1/1/1	The translation fails to preserve the original meaning and produces an incomprehensible sentence unrelated to the source content.
Mempertinggi semangat juang orang lain, menjatuhkan kewibawaan diri sendiri.	长别人志气，灭自己威风。	3/3/3	The translation adopts a concise and idiomatic Chinese expression with strong stylistic naturalness.

Figure 6: Representative annotation examples from the AFE dataset. The examples illustrate that adequacy, fluency, and elegance capture different aspects of translation quality, particularly for literal and idiomatic translations.

B Preference Dataset Construction

B.1 Translation Generation Prompt

To construct rejected responses for preference learning, we use DeepSeek-v4-pro to generate alternative Indonesian–Chinese translations from the same source sentences. Large language models trained on large-scale Chinese corpora can typically produce fluent and semantically understandable translations while still exhibiting literal or stylistically rigid expressions in low-resource translation scenarios. This property makes them suitable for constructing stylistically weaker candidate translations for preference learning.

The following prompt is used during translation generation:

Instruction Prompt

> You are an Indonesian–Chinese translation expert. Translate the given Indonesian text into natural and fluent Simplified Chinese while preserving the original meaning. Provide only the Chinese translation without additional explanation.

The input and output formats are defined as follows.

Input Format

> [{"input": "<formal Indonesian text>"}]

Output Format

> [{ "input": "<formal Indonesian text>", "output": "<Chinese translation>" }]

The model is instructed not to generate explanations, notes, Markdown formatting, or additional fields beyond the translation output.

We use the generated translations as candidate rejected responses for preference construction. Since our goal is to model stylistic naturalness rather than semantic correctness alone, we do not manually enforce literal translation behavior during prompting.

B.2 Preference Filtering Strategy

After generation, all candidate translations are manually filtered before inclusion in the preference dataset. We remove samples containing severe semantic errors, hallucinations, incomplete translations, or unintelligible outputs. We retain only preference pairs where the chosen and rejected translations remain broadly comparable in adequacy and basic fluency, while differing primarily in stylistic naturalness, idiomaticity, or expression quality.

This filtering process encourages the reward model to learn subtle stylistic distinctions instead of relying on trivial semantic mismatches.

B.3 Additional Preference Examples

Figure 7 presents additional preference examples from our stylistic preference dataset. The examples include both idiomatic expressions and general sentence translation cases, illustrating that the chosen and rejected translations are typically comparable in semantic correctness while differing mainly in stylistic naturalness, idiomaticity, and expression quality.

Additional Preference Examples

Source Sentence	Chosen Translation	Rejected Translation	Main Difference
Tempat tinggal bangsawan tak gampang dimasuki orang awam.	侯门深似海	贵族的居住地方不容易被普通的人进入。	Literal translation vs idiomatic Chinese expression
Orang datang melampaui orang asal.	后来者居上	人来超过原来的人。	Weak idiomaticity and unnatural wording
Bapak harimau takkan punya anak anjing.	虎父无犬子	老虎父亲不会有孩子狗。	Translationese and broken expression
Bila enam orang tidak rukun karena berselisih pendapat, musuh yang sudah tertangkap juga bisa melarikan diri.	六个人闹意见的话，抓到手的敌人也会逃跑	如果六个人不和睦因为意见的分歧，已经捉住的敌人也可以自己逃跑。	Semantically acceptable but stylistically rigid
Gadis yang keburu kawin, tak kan menunjukkan perhatian kepada suaminya; pelancong yang berjalan dengan memburu waktu, mudah tersesat.	急于出嫁的姑娘，对丈夫不体贴；忙于赶路的旅客，容易迷失方向	急着结婚的少女，不会对她的丈夫表示注意；追逐时间行走的游客，容易走失。	Natural human-like expression vs literal rendering

Figure 7: Additional preference examples from the stylistic preference dataset. The chosen translations exhibit more natural and idiomatic Chinese expression, while the rejected translations remain semantically understandable but stylistically literal or rigid.

C Prompt of Transjudge

/* Task prompt */
Evaluate the quality of the given Chinese translation of an Indonesian source sentence. Focus on whether the translation preserves the source meaning, reads naturally in Chinese, and achieves appropriate wording and style.
Adequacy
1 point: Poor translation, contains significant mistranslation, omission, addition, semantic distortion, or semantic reversal.
2 points: Fair translation, preserves the main meaning but contains minor inaccuracies or semantic deviations.
3 points: Good translation, accurately and completely preserves the meaning of the source sentence.
Fluency
1 point: Poor translation, unnatural, incoherent, or grammatically incorrect, making it difficult to understand.
2 points: Fair translation, understandable but awkward, partially unnatural, or mildly ungrammatical.
3 points: Good translation, fluent, coherent, natural, and consistent with Chinese expression norms.
Elegance
1 point: Poor translation, stiff, overly literal, stylistically inappropriate, or culturally weak.
2 points: Fair translation, acceptable but plain, partially literal, or insufficiently refined.
3 points: Good translation, idiomatic, stylistically appropriate, culturally natural, and able to preserve the rhetorical or pragmatic effect of the source when applicable.

/* Input Data */:

> Indonesian source sentence: {source}
> Chinese translation: {translation}
> Evaluation:

/* Output Text */:

> { “adequacy”: {score},
> “fluency”: {score},
> “elegance”: {score} }

D LLM Evaluator Settings

For fair comparison, all LLM-based evaluators use the same evaluation prompt described in Appendix C. We evaluate the following proprietary and open-source judge models through their official APIs or official inference interfaces:

- **GPT-4o**: `gpt-4o`.
- **DeepSeek**: `deepseek-v4-pro`.
- **Qwen**: `Qwen3.6-Plus`.

All API-based evaluators are run with temperature set to 0 and a maximum generation length of 4096 tokens. We use deterministic decoding to reduce variance in the scoring outputs. All model responses are parsed as JSON arrays following the required AFE scoring format.

E Human Evaluation Details

This appendix provides additional details about the human evaluation protocol and the reliability of the human annotations. The main paper reports the agreement between automatic evaluators and human judgments; here, we focus on how the human reference judgments were collected and whether the annotators produced a stable reference signal.

E.1 Annotation Protocol

We randomly sample 100 Indonesian–Chinese translation pairs for human evaluation. Each example contains an Indonesian source sentence and its corresponding Chinese translation. The examples are shuffled before annotation so that annotators do not see the original data order. Annotators follow the AFE scoring rubric described in Appendix A; we do not repeat the rubric here.

The analysis uses annotations from three independent annotators. For each example, we compute a human consensus score by taking the median of the three annotators' scores for each AFE dimension. The overall score is then computed as the sum of faithfulness, fluency, and elegance, and therefore ranges from 3 to 9. Items with non-numeric annotation artifacts are excluded only from the affected calculations.

E.2 Inter-Annotator Consistency

We report inter-annotator consistency to quantify the reliability of the human judgments themselves. Table 6 shows average pairwise correlations and agreement statistics among the three annotators. The annotators show a stable agreement pattern on the overall score, with an average Pearson correlation of 0.74 and an average Kendall correlation of 0.56. Dimension-level agreement is strongest for fluency and relatively lower for elegance, which is expected because stylistic preference is more subjective than adequacy or grammatical fluency. The high within-one-point agreement across all dimensions indicates that most disagreements are minor.

These results indicate that the human annotation protocol yields a reasonably stable reference signal. At the same time, the relatively lower agreement on elegance confirms that stylistic quality is the most subjective dimension, which motivates reporting dimension-level results rather than relying only on a single overall score.

Table 6: Inter-annotator consistency averaged over annotator pairs.

Metric	r	ρ	τ	Exact	Within-1	MAD
Overall	0.74	0.65	0.56	45.31	74.83	0.99
Faithfulness	0.67	0.64	0.58	65.45	97.65	0.37
Fluency	0.71	0.64	0.60	71.33	97.33	0.31
Elegance	0.62	0.59	0.55	64.67	94.00	0.41

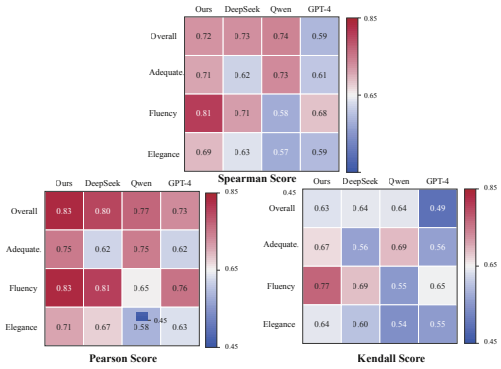

Figure 8: Heatmap of correlations between automatic evaluators and human consensus across the overall score and the three evaluation dimensions. Darker colors indicate stronger agreement with human consensus.

E.3 Additional Visualization of Evaluator-Human Agreement

Figure 8 visualizes the correlation patterns between automatic evaluators and human consensus across the overall score and the three AFE dimensions.