# BEYONDADEEVAL: MACHINE TRANSLATION EVALUATION BEYOND SEMANTIC ADEQUACY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Machine translation metrics and LLM-based evaluators have made substantial progress in translation evaluation. However, existing methods such as COMET primarily focus on translation correctness and offer limited interpretability. Evaluating overall translation quality beyond semantic adequacy remains challenging. We present BeyondAdeEval (BAE), a reference-free and interpretable method for evaluating translation quality from semantic adequacy to finer-grained quality beyond semantic adequacy. We design **BAE-RUBRIC**, a principled, human-aligned hierarchical rubric that formalizes translation quality, guides data construction, and enables our model to learn structured, rubric-grounded rationales. Building on this rubric, we introduce (i) **BAE-BENCH**, a preference-graded translation benchmark constructed through controlled modification of high-quality translations, with 6,960 training samples and a human-annotated test set supporting pointwise scoring and pairwise ranking; and (ii) **BAE-SERIES**, a family of evaluation models from 4B to 9B parameters that produces human-aligned, multidimensional, and explainable translation evaluations. On BAE-BENCH, our models outperform all evaluated baselines, achieving relative improvements of 28% and 24% in Spearman and Pearson correlation over the strongest learned metrics. We also achieve state-of-the-art performance on T-index (Human) and demonstrate zero-shot cross-lingual generalization on MQM24. We will release our code and data upon acceptance.

## 1 INTRODUCTION

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in machine translation (MT), substantially raising the adequacy and fluency of generated translations (Xu et al., 2024a;b; 2025; Yang et al., 2025). However, high-quality translation remains difficult to evaluate and optimize (Popov et al., 2025). Two translations can be equally accurate while differing substantially in dimensions such as rhythm, idiomaticity, and register. Such differences are captured by the broader criteria of expressiveness and elegance, which shape how a target-language reader reconstructs character, atmosphere and pacing.

Conventional MT metrics such as BLEU, BLEURT, and COMET primarily assess semantic adequacy or overall translation quality through lexical overlap or learned correlations with human judgments (Papineni et al., 2002; Sellam et al., 2020; Rei et al., 2020; Juraska et al., 2024). More recent learned and LLM-based evaluators enable more fine-grained and interpretable assessment, often by adopting structured evaluation frameworks such as Multidimensional Quality Metrics(MQM) to identify errors across multiple quality dimensions and aggregate their types and severities into an overall quality judgment (Guerreiro et al., 2024; Kim, 2025; Maheswaran et al., 2025; Kocmi & Federmann, 2023; Fernandes et al., 2023; Lu et al., 2024; Sproat et al., 2025; Finkelstein et al., 2025). A growing line of work has further broadened translation evaluation beyond semantic adequacy (Zhang et al., 2025; Whitehouse et al., 2026). T-index, for example, explicitly targets translationese (Liu et al., 2025), while FiRE evaluates translations along faithfulness, fluency, and stylistic consistency before aggregating these criterion-level judgments into an overall preference (Gao et al., 2026).

Despite this progress, reliably distinguishing translations that are semantically correct yet differ substantially in quality beyond adequacy remains challenging. Conventional regression-based metrics



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

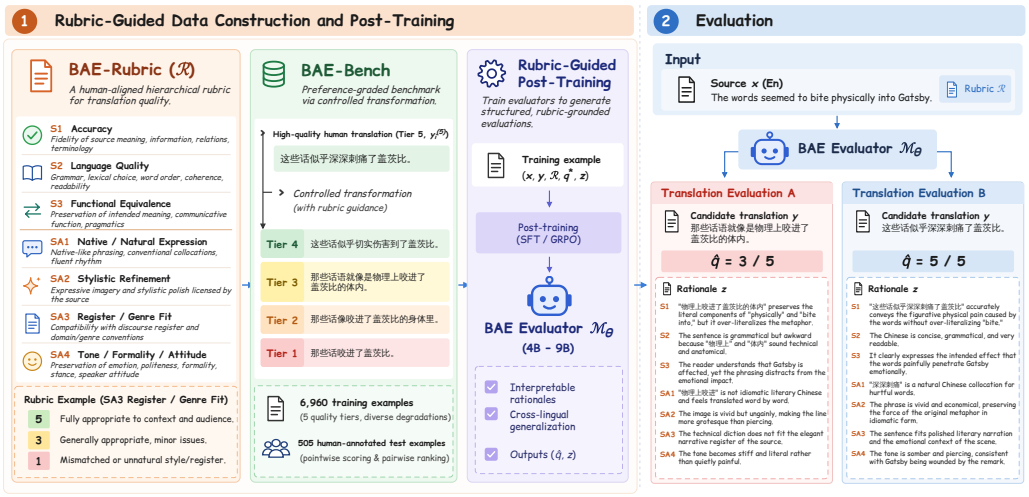



Figure 1: Overview of BAE.

compress translation quality into a single score and provide little explicit characterization of dimensions such as naturalness, stylistic appropriateness, and overall writing quality (Popov et al., 2025; Flamich et al., 2025). More structured LLM-based evaluators broaden the scope of evaluation through multidimensional criteria or error taxonomies, but these formulations are often oriented toward identifying errors rather than differentiating subtle tiers among otherwise correct translations. Moreover, prompting-based approaches largely rely on the evaluation capabilities already encoded in general-purpose LLMs, leaving open how reliably prompting alone can elicit and calibrate these capabilities for fine-grained quality discrimination. These limitations motivate a more structured and discriminative formulation of translation quality that covers both semantic adequacy and finer-grained qualities beyond semantic adequacy, while explicitly distinguishing between the two.

To address this gap, we introduce BAE, a reference-free and interpretable framework for evaluating translation quality across the full quality spectrum, from semantic adequacy to finer-grained qualities beyond semantic adequacy. We design BAE-RUBRIC, a human-aligned hierarchical rubric spanning three core correctness dimensions and four dimensions beyond semantic adequacy, and use it to construct BAE-BENCH, an English-to-Chinese paragraph-level benchmark with preference-graded translations derived from professional human translations through controlled modification. Based on this benchmark, we train BAE-SERIES, a family of 4B–9B evaluators that jointly predict translation quality and generate rubric-grounded explanations. Our best model achieves a Spearman correlation of 0.802 with expert judgments on BAE-BENCH, outperforming the strongest evaluated baseline of 0.764. Despite being trained only on English-to-Chinese pointwise evaluation, BAE-SERIES generalizes effectively to external pairwise benchmarks: BAE-9B reaches 77.3% agreement on the human-annotated T-index benchmark (Liu et al., 2025), surpassing T-index by 1.3 percentage points, and achieves 72.2% on MQM24 En–De (Freitag et al., 2024). The latter result is particularly notable because the model receives no German training data or pairwise supervision, demonstrating transfer across both evaluation settings and languages.

Our main contributions are:

(i) **BAE-RUBRIC.** We introduce a human-aligned hierarchical rubric covering both core correctness and finer-grained quality beyond semantic adequacy.

(ii) **BAE-BENCH.** We construct a preference-graded paragraph-level benchmark with five ordered quality tiers, comprising 6,960 training examples and a human-annotated test set of 505 candidates, supporting both pointwise scoring and pairwise ranking.

(iii) **BAE-SERIES.** We train a family of reference-free evaluators that produce quality scores and rubric-grounded rationales, achieving strong human alignment and generalization across evaluation settings and languages. BAE-SERIES evaluators may serve as a reward model for further translation optimization.

Table 1: Overview of BAE-RUBRIC.

| Dimension | Evaluation focus | Sub-metrics |
|---|---|---|
| S1 Accuracy | Fidelity of source meaning, information, relations, terminology | Mistranslation • Omission • Addition • Misrepresented Relations • Terminology Misuse |
| S2 Language Quality | Grammar, lexical choice, word order, coherence, readability | Basic Grammar Correctness • Lexical Appropriateness • Natural Word Order • Coherence & Readability |
| S3 Functional Equivalence | Preservation of intended meaning, communicative function, pragmatics | Clarity of Intended Meaning • Communicative Function Preservation • Pragmatic Interpretability |
| SA1 Native / Natural Expression | Native-like phrasing, conventional collocations, fluent rhythm | Native-like Expression • Natural Collocations • Smooth Rhythm & Flow |
| SA2 Stylistic Refinement | Expressive imagery and stylistic polish licensed by the source | Vivid Imagery • Stylistic Refinement |
| SA3 Register / Genre Fit | Compatibility with discourse register and domain/genre conventions | Fit with Discourse Register • Domain/Genre Suitability |
| SA4 Tone / Formality / Attitude | Preservation of emotion, politeness, formality, stance, speaker attitude | Preserved Emotional Tone • Proper Level of Formality • Attitude & Intent Accuracy |

# 2 METHOD

## 2.1 PRELIMINARY

We formulate translation evaluation as

$$\tau = \mathcal{M}_\theta(x, y; \mathcal{R}) = (\hat{q}, z), \qquad \hat{q} \in \{1, \ldots, 5\},$$

where $x$ denotes a source text segment consisting of 2–3 sentences, $y$ denotes its candidate translation, and $\mathcal{R}$ denotes the evaluation rubric. The evaluator model $\mathcal{M}_\theta$ outputs an overall translation-quality score $\hat{q}$ together with a rubric-grounded rationale $z$ that explains the assigned score. For supervised training, we denote the target overall quality score by $q^*$.

To construct preference-graded training data, we begin with a set of Tier-5 source–translation pairs,

$$\mathcal{D}_5 = \left\{ \left( x_i, y_i^{(5)} \right) \right\}_{i=1}^{N},$$

which serve as high-quality anchors for constructing translations at lower quality tiers. Unlike reference-based MT metrics, BAE requires only the source text $x$ and candidate translation $y$ at inference time, without access to a reference translation. The complete data-construction and post-training pipeline is summarized in Algorithm 1.

## 2.2 BAE-RUBRIC

To capture translation quality across both semantic adequacy and finer-grained quality beyond semantic adequacy, we introduce BAE-RUBRIC, a hierarchical rubric built around seven dimensions of translation quality. Table 1 summarizes their operational scope. Our rubric draws on prior evaluation frameworks covering structured error categories, source fidelity, target-language quality, and higher-order properties such as fluency and style (Lommel et al., 2014; Fernandes et al., 2023; Gao et al., 2026; Kim, 2025). Based on these principles, we organize translation quality into three core dimensions (accuracy, language quality, and functional equivalence) and four dimensions beyond semantic adequacy (native expression, stylistic refinement, register/genre fit, and tone/formality/attitude).

The initial rubric was refined through a small-scale pilot annotation with expert annotators. Professional annotators were asked to apply the guidelines and flag ambiguous criteria, unclear boundaries, and cases requiring further clarification; their feedback was incorporated into subsequent revisions. The resulting rubric guides both data construction and rationale supervision, as described in Sections 2.3 and 2.4.

Table 2: Statistics of the BAE dataset and BAE-BENCH.

| Split | Source segments | Translation candidates | Avg. src. words | Avg. tgt. chars | Avg. rat. words | Pairwise pairs |
|---|---|---|---|---|---|---|
| Train | 1,392 | 6,960 | 27.3 | 46.5 | 134.2 | 13,920 |
| Test | 101 | 505 | 43.3 | 68.0 | — | 1,010 |

## 2.3 DATA CONSTRUCTION

### 2.3.1 TRAIN-SET CONSTRUCTION

**Tier-5 Anchor Curation.** We first curate source segments from published novels together with their professional translations, excluding segments that are too short to exhibit meaningful quality variation. Each source–translation pair is then verified against BAE-RUBRIC: an LLM flags potential issues for human inspection, and annotators correct or discard pairs that do not meet Tier-5 quality. The resulting verified pairs form the Tier-5 anchor set $\mathcal{D}_5$.

**Lower-level Data Curation.** Given a verified Tier-5 anchor $(x, y^{(5)}) \in \mathcal{D}_5$, we construct lower-quality variants by applying rubric-guided, level-specific modifications to $y^{(5)}$. Conditioned on $(x, y^{(5)}, \mathcal{R})$, the transformation model $\mathcal{G}_d$ edits the anchor under explicit constraints on which rubric dimensions may change and by how much, producing $\{y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}\}$.

We use controlled modification rather than outputs from heterogeneous MT systems because naturally occurring system outputs do not reliably cover the full quality spectrum or adjacent tiers for the same source. This construction yields source-matched groups $\mathcal{Y}(x) = \{y^{(1)}, \ldots, y^{(5)}\}$ with the intended ordering $y^{(1)} \prec \cdots \prec y^{(5)}$. Human annotators inspected a subset of the groups to verify that adjacent levels remain distinguishable and the intended ordering is preserved. Figure 1 shows an example, with full construction details in Appendix B.

**Rationale Construction.** A separate annotation teacher $\mathcal{G}_a$ generates a rubric-grounded rationale $\tilde{z}$ conditioned on $(x, y^{(q)}, q, \mathcal{R})$. We provide the assigned quality tier $q$ explicitly so that rationale generation focuses on explaining the supervised tier under BAE-RUBRIC, rather than independently re-estimating the score. The teacher model version is given in Appendix E.1, and the rationale-generation procedure is detailed in Appendix B.

### 2.3.2 TEST-SET CONSTRUCTION

We construct a source-disjoint human-annotated test set of BAE-BENCH comprising 505 candidates from 101 held-out source groups. Annotators grade each candidate independently under BAE-RUBRIC and assign a single overall quality score, without access to its construction tier or the intended within-group ordering. These independent human grades, rather than the construction tiers, serve as the gold labels for test-set evaluation. Annotation guidelines and annotator qualifications are provided in Appendix D.

## 2.4 RUBRIC-GUIDED POST-TRAINING

### 2.4.1 SUPERVISED FINE-TUNING

We apply LoRA (Hu et al., 2022) fine-tuning to adapt the base model to BAE-RUBRIC and the structured evaluation format. In the supervised stage, each example pairs the source passage, candidate translation, and rubric with a target response $\tau^*$ containing the overall quality label and the teacher-generated rationale. The model is trained autoregressively over the complete response:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y,q^*,\tilde{z})\sim\mathcal{D}_{\mathrm{SFT}}}\left[\sum_{t=1}^{|\tau^*|} w_t \log p_\theta(\tau_t^* \mid x, y, \mathcal{R}, \tau_{<t}^*)\right]. \tag{1}$$

Here, $\tau_t^*$ is the $t$-th token of the serialized target response, and $w_t$ is its loss weight; score-token weighting is specified in Table 5. This stage adapts the model to produce the required 1–5 quality score together with a rubric-grounded rationale.

#### 2.4.2 REINFORCEMENT LEARNING

We further fine-tune the evaluator with GRPO (Shao et al., 2024), using the overall quality label as the outcome supervision signal. Our main design choice in this stage is the reward, which reflects the ordinal tiers defined by BAE-RUBRIC.

**Reward.** We use a discrete ordinal reward that decreases with the distance between the predicted and target tiers. For the $j$-th sampled evaluation trajectory, the reward is

$$R_j = \begin{cases} 1, & |\hat{q}_j - q^*| = 0, \\ 0.4, & |\hat{q}_j - q^*| = 1, \\ -0.2, & |\hat{q}_j - q^*| = 2, \\ -0.6, & |\hat{q}_j - q^*| \geq 3, \\ -2, & \text{if the response cannot be parsed.} \end{cases} \tag{2}$$

The first four cases apply to parseable responses, with $\hat{q}_j$ denoting the predicted overall score and $q^*$ the target quality label. The reward encodes the ordered distance between levels on the 1–5 quality scale.

**Training paradigm.** For each training instance, the current policy samples a group of $G$ complete evaluation trajectories and computes their relative advantages:

$$\begin{aligned} \tau_j = (\hat{q}_j, z_j) \sim p_\theta(\cdot \mid x, y, \mathcal{R}), \qquad j = 1, \ldots, G, \\ A_j = \frac{R_j - \text{mean}(R_1, \ldots, R_G)}{\text{std}(R_1, \ldots, R_G) + \epsilon}. \end{aligned} \tag{3}$$

Here, $\epsilon$ is a small constant for numerical stability. GRPO normalizes rewards among trajectories generated for the same input and optimizes the policy toward trajectories with larger relative advantages. Following the post-training procedure, we train a family of evaluation models, BAE-SERIES, with parameter counts ranging from 4B to 9B. Each model is initialized from the corresponding Qwen3.5 model (Qwen Team, 2026) and trained on the same data and recipe. The resulting evaluator performs reference-free translation evaluation by taking $(x, y, \mathcal{R})$ as input and producing an overall score $\hat{q}$ and a rubric-grounded rationale $z$. Full implementation and training details are provided in Appendix A.

## 3 EXPERIMENTS

### 3.1 BASELINES

We compare against four families of reference-free MT evaluators. **Regression-based** methods include COMET-QE-wmt20 (Rei et al., 2020), Kiwi-XXL (Rei et al., 2023), xCOMET-XXL (Guerreiro et al., 2024), and MetricX-24-XXL (Juraska et al., 2024). **Ranking- and likelihood-based** methods include MT-Ranker-XXL (Moosa et al., 2024), FiRE (Gao et al., 2026), and T-index (Liu et al., 2025). **Error-based** methods include GEMBA-MQM (Kocmi & Federmann, 2023), Rubric-MQM (Kim, 2025), EAPrompt (Lu et al., 2024), and AutoMQM (Fernandes et al., 2023). We additionally evaluate **general-purpose LLMs**, including Qwen3.5-9B (Qwen Team, 2026), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), and GPT-6 Astra (OpenAI, 2026), both with their standard evaluation prompt and, where applicable, with BAE-RUBRIC. FiRE and the error-based prompting methods use Qwen3.5-9B as the backbone. Model versions and implementation details are provided in Appendix E.1.

### 3.2 EVALUATION SETUP

We evaluate BAE-4B and BAE-9B, initialized from the corresponding Qwen3.5 backbones, and use BAE-9B for detailed analyses unless otherwise specified. On BAE-BENCH, each candidate is evaluated independently from its source text and BAE-RUBRIC, producing a 1–5 quality score and a rubric-grounded rationale. We additionally evaluate high-quality discrimination using within-source tier-4 vs. tier-5 pairs, where ties are counted as incorrect. For external evaluation, we use the human

Table 3: Agreement with human judgments on BAE-BENCH. Bold indicates the best result in each column.

| Evaluator | Spearman ↑ | Pearson ↑ | MAE ↓ | Score Acc. ↑ |
|---|---|---|---|---|
| **Regression-Based** | | | | |
| COMET-QE-wmt20 | 0.565 | 0.581 | 0.974 | — |
| Kiwi-XXL | 0.606 | 0.635 | 0.905 | — |
| xCOMET-XXL | 0.464 | 0.465 | 1.089 | — |
| MetricX-24-XXL | 0.628 | 0.642 | 0.891 | — |
| **Error-Based** | | | | |
| GEMBA-MQM | 0.527 | 0.529 | 1.006 | — |
| Rubric-MQM | 0.492 | 0.470 | 1.044 | — |
| EAPrompt | 0.402 | 0.368 | 1.137 | — |
| AutoMQM | 0.445 | 0.448 | 1.141 | — |
| **LLM-Based** | | | | |
| Qwen3.5-9B | 0.666 | 0.679 | 1.085 | 22.0 |
| DeepSeek-V4.1-Flash | 0.701 | 0.707 | 0.816 | 33.5 |
| GPT-6 Astra | 0.764 | 0.771 | 0.836 | 37.4 |
| **BAE-Rubric-Augmented LLMs** | | | | |
| Qwen3.5-9B + BAE-RUBRIC | 0.683 | 0.686 | 0.768 | 40.6 |
| DeepSeek-V4.1-Flash + BAE-RUBRIC | 0.718 | 0.721 | 0.679 | 45.7 |
| GPT-6 Astra + BAE-RUBRIC | 0.795 | **0.801** | 0.582 | 51.1 |
| **BAE-SERIES** | | | | |
| **BAE-4B** | **0.802** | 0.796 | **0.566** | **55.4** |
| **BAE-9B** | 0.776 | 0.763 | 0.620 | 53.3 |

pairwise judgments from T-index (Human) (Liu et al., 2025) and MQM24 En–De (Freitag et al., 2024); ties are again counted as incorrect. All BAE training data are English–Chinese, so MQM24 En–De evaluates zero-shot cross-lingual transfer. Training and evaluation prompts are provided in Appendices A and F.

### 3.3 METRICS

On BAE-BENCH, we report Spearman and Pearson correlation with human 1–5 scores, MAE, and exact score accuracy. For evaluators with native continuous outputs, correlations are computed directly and scores are quantile-mapped to the 1–5 scale only for MAE; exact score accuracy is reported only for evaluators that directly predict discrete quality labels. For pairwise evaluation, continuous or pointwise scores are converted to preferences by comparing the two candidates, while ranking-based methods use their native pairwise predictions. T-index (Human), MQM24 En–De, and the tier-4-vs-tier-5 diagnostic are reported using pairwise accuracy.

### 3.4 PERFORMANCE ON BAE-BENCH

**Overall pointwise evaluation.** Table 3 shows that BAE-SERIES substantially improves alignment with human quality judgments over conventional learned metrics and general-purpose LLM evaluators. BAE-4B achieves the highest Spearman correlation (0.802), lowest MAE (0.566), and highest exact score accuracy (55.4%) among the evaluated methods, while BAE-9B reaches 0.776 Spearman and 0.763 Pearson. Compared with MetricX-24-XXL, the strongest regression-based baseline on BAE-BENCH, BAE-4B improves Spearman from 0.628 to 0.802 and Pearson from 0.642 to 0.796.

Importantly, BAE-RUBRIC already improves general-purpose LLM evaluators without any post-training. Providing BAE-RUBRIC at inference time consistently improves Qwen3.5-9B, DeepSeek-V4.1-Flash, and GPT-6 Astra. The gains are particularly pronounced in absolute score calibration: for example, GPT-6 Astra improves from 0.836 to 0.582 MAE and from 37.4% to 51.1% exact score accuracy. This suggests that explicitly decomposing translation quality helps general-purpose LLMs map their judgments more consistently onto the human 1–5 quality scale.

Post-training provides a further substantial improvement beyond rubric prompting alone. Starting from Qwen3.5-9B+BAE-RUBRIC, supervised fine-tuning raises Spearman from 0.683 to 0.776,

reduces MAE from 0.768 to 0.621, and improves exact score accuracy from 40.6% to 53.1%. These results indicate that learning from rubric-defined quality gradients is substantially more effective than eliciting the same criteria through prompting alone.

**Discriminating high-quality translations.** Aggregate score correlation can be dominated by relatively easy distinctions involving semantic or grammatical errors. We therefore separately evaluate the most challenging region of the quality spectrum: tier-4 versus tier-5 translations of the same source. These candidates are both high-quality translations, with the distinction primarily arising from finer-grained properties such as native expression, stylistic refinement, register, and tone. This setting therefore directly tests the beyond-correctness discrimination targeted by BAE.

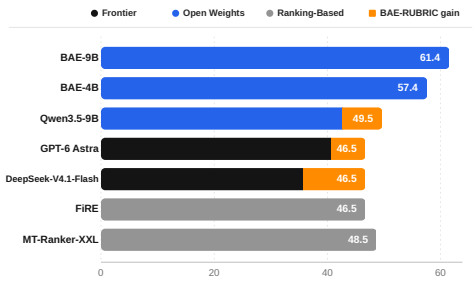



Figure 2: Pairwise accuracy (%) on the tier-4-vs-tier-5 pairs in BAE-BENCH.

As shown in Figure 2, BAE-9B correctly orders 61.4% of these pairs, substantially outperforming the same Qwen3.5-9B backbone (42.6%) and Qwen3.5-9B+BAE-RUBRIC (49.5%), as well as the other evaluated baselines. The substantially larger gain in this high-quality regime indicates that post-training is particularly useful when semantic adequacy alone is insufficient to distinguish competing translations.

**Effect of model scale.** Interestingly, BAE-4B performs better than BAE-9B on the in-domain BAE-BENCH test set, whereas the ordering reverses on the external benchmarks discussed below. One possible explanation is an adaptation–generalization trade-off: given the relatively small 6,960-example training set, the smaller model may adapt more strongly to the in-domain quality distribution, while the larger backbone retains stronger pretrained capabilities under distribution shift. Although this interpretation is not conclusive, the external results below consistently favor BAE-9B, suggesting that model scale becomes more important when transferring beyond the training setting.

### 3.5 GENERALIZATION TO EXTERNAL HUMAN BENCHMARKS

The controlled construction of BAE-BENCH provides dense supervision across adjacent quality levels, but also raises the question of whether the learned evaluation capability transfers beyond this construction. We therefore evaluate on two independent human-annotated benchmarks with different candidate sources and annotation protocols: T-index (Human), which targets translationese and native-like quality, and MQM24 En–De, which further tests zero-shot transfer to an unseen language pair.

**Beyond-adequacy transfer on T-index.** T-index (Human) is particularly relevant to our setting because it evaluates translationese, a form of unnatural target-language expression closely related to the dimensions beyond semantic adequacy captured by BAE-RUBRIC (Liu et al., 2025). Unlike BAE-BENCH, its candidates are naturally occurring outputs from multiple MT systems and LLMs rather than translations constructed at predefined quality levels, and its pairwise preferences are independently annotated by humans. As shown in Figure 3, despite being trained only on pointwise English–Chinese supervision, BAE-9B reaches 77.3% agreement with human preferences, improving substantially over the same Qwen3.5-9B backbone (66.0%) and exceeding the original T-index result (76.0%). This provides direct evidence that the learned quality distinctions transfer beyond the controlled construction of BAE-BENCH.

**Agreement under human uncertainty.** The remaining errors are concentrated in comparisons that are also difficult for human annotators. On T-index, BAE-9B agrees with the majority on 66.7% of the 45 pairs with three-of-five annotator agreement, 78.8% of the 66 pairs with four-of-five agreement, and 87.2% of the 39 unanimously judged pairs. The monotonic increase with human consensus suggests that model errors are concentrated in intrinsically ambiguous cases rather than being uniformly distributed across the benchmark.

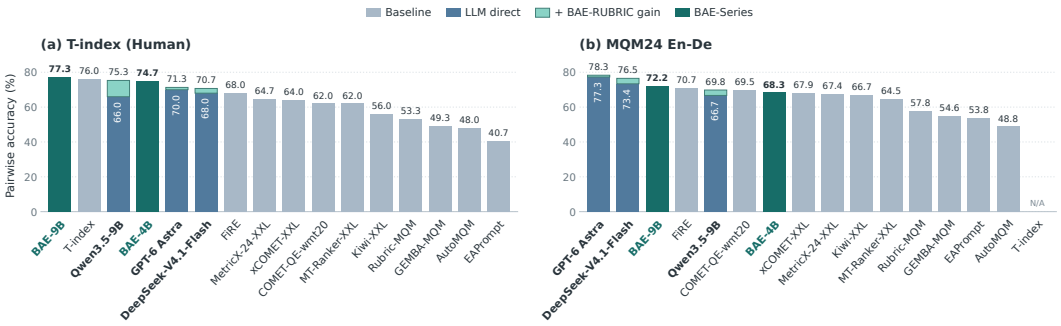



Figure 3: Pairwise accuracy (%) on T-index (Human) and MQM24 En–De.

Table 4: Effects of BAE-RUBRIC and post-training on BAE-BENCH, T-index (Human), and MQM24 En–De. Accuracies are percentages; bold indicates the best result in each column.

| Setting | BAE-Bench | | | | | T-index (Human) | MQM24 En–De |
|---|---|---|---|---|---|---|---|
| | Spearman ↑ | Pearson ↑ | MAE ↓ | Score Acc. ↑ | 4-vs-5 Acc. ↑ | Pair Acc. ↑ | Pair Acc. ↑ |
| Qwen3.5-9B | 0.666 | 0.679 | 1.085 | 22.0 | 42.6 | 66.0 | 66.7 |
| Qwen3.5-9B + BAE-RUBRIC | 0.683 | 0.686 | 0.768 | 40.6 | 49.5 | 75.3 | 69.8 |
| BAE-9B (SFT) | **0.776** | **0.764** | 0.621 | 53.1 | **61.4** | 75.3 | 71.8 |
| BAE-9B (SFT + GRPO) | **0.776** | 0.763 | **0.620** | **53.3** | **61.4** | **77.3** | **72.2** |

**Cross-lingual transfer on MQM24.** MQM24 En–De provides a complementary test under a conventional MT-evaluation setting with human MQM judgments (Freitag et al., 2024). Since BAE-SERIES is trained exclusively on English–Chinese pointwise evaluation, this setting requires transfer across both language pair and evaluation protocol. As shown in Figure 3, BAE-9B achieves 72.2% pairwise accuracy, outperforming the specialized regression-, ranking-, and error-based evaluators shown in the figure, including FiRE (70.7%) and MetricX-24-XXL (67.4%). It also improves by 5.5 points over its unadapted Qwen3.5-9B backbone, indicating that the learned evaluation criteria are not specific to English–Chinese.

The rubric itself also exhibits strong transfer. Adding BAE-RUBRIC improves Qwen3.5-9B from 66.7% to 69.8% on MQM24, DeepSeek-V4.1-Flash from 73.4% to 76.5%, and GPT-6 Astra from 77.3% to 78.3%, the strongest result in Figure 3. Together with the T-index results, these findings suggest that BAE-RUBRIC captures translation-quality criteria that generalize across data-construction procedures, evaluation protocols, model families, and languages.

### 3.6 ABLATION STUDY

To understand where the gains of BAE arise, we ablate inference-time rubric guidance, supervised fine-tuning, and GRPO. Table 4 shows a clear progression across the three stages.

First, providing BAE-RUBRIC at inference time consistently improves the Qwen3.5-9B backbone. The largest gains appear in absolute score calibration on BAE-BENCH, where MAE decreases from 1.085 to 0.768 and exact score accuracy increases from 22.0% to 40.6%; T-index pairwise accuracy also rises from 66.0% to 75.3%. Thus, an explicit multidimensional rubric already provides a strong evaluation prior without parameter updates.

Second, SFT contributes the largest improvement on BAE-BENCH. Relative to Qwen3.5-9B+BAE-RUBRIC, SFT raises Spearman from 0.683 to 0.776, lowers MAE from 0.768 to 0.621, and increases tier-4-vs-tier-5 accuracy from 49.5% to 61.4%. These gains show that learning from rubric-defined quality gradients provides capabilities that cannot be obtained through inference-time prompting alone.

Finally, GRPO leaves in-domain correlation and high-quality discrimination largely unchanged, but improves external preference agreement: T-index accuracy increases from 75.3% to 77.3%, and MQM24 En–De from 71.8% to 72.2%. This suggests that the ordinal outcome objective primarily refines preference ordering and improves transfer, while most of the in-domain quality calibration is learned during SFT.

# 4 Related Work

## 4.1 Machine Translation and Learned Evaluation

Machine translation has increasingly shifted toward large language models, with translation-specific fine-tuning, preference optimization, and context-aware or agentic systems further improving translation quality (Lim et al., 2025; Tanzer et al., 2024; Aycock et al., 2025; Xu et al., 2024a;b; 2025; Wu et al., 2025; Wang et al., 2025; Lu et al., 2025). As modern systems approach high semantic adequacy, evaluation must increasingly distinguish translations that are similarly correct but differ in naturalness, style, and contextual coherence.

Traditional metrics such as BLEU rely on reference overlap (Papineni et al., 2002), while learned metrics including BLEURT, COMET, xCOMET, and MetricX-24 improve alignment with human judgments through pretrained representations and richer supervision (Sellam et al., 2020; Rei et al., 2020; Guerreiro et al., 2024; Juraska et al., 2024). Despite strong aggregate correlation, these methods primarily produce scalar quality estimates or error annotations and provide limited supervision for discriminating fine-grained quality differences among high-quality translations.

## 4.2 Human-Aligned Evaluation Beyond Correctness

Recent work broadens MT evaluation through structured criteria and dimensions beyond semantic adequacy. MQM provides a widely used framework for categorizing translation errors (Lommel et al., 2024), while methods such as GEMBA-MQM and related LLM-based evaluators operationalize structured error analysis through prompting or learned models (Kocmi & Federmann, 2023; Sproat et al., 2025; Popov et al., 2025). Other approaches explicitly target higher-order quality: T-index measures translationese (Liu et al., 2025), MENLO and DiscoX study native-like and discourse-level quality (Whitehouse et al., 2026; Zhao et al., 2026), and literary-translation evaluation considers linguistic, stylistic, and reader-centered properties (Shafayat et al., 2024; Zhang et al., 2025; Kim et al., 2025; Ferstler et al., 2026).

Pairwise methods provide another direction: FiRE compares translations across faithfulness, fluency, and stylistic consistency (Gao et al., 2026), while ReMedy learns evaluators from human preferences (Tan & Monz, 2025). In contrast, BAE trains a dedicated evaluator on rubric-defined quality gradients, combining absolute graded assessment with explicit evaluation of both correctness and finer-grained quality beyond correctness.

# 5 Limitations and Future Directions

Our current training data focus on English–Chinese literary translation and use controlled modifications of professional translations to construct quality gradients. Although the strong transfer to T-index and MQM24 suggests that BAE generalizes beyond this setting, broader coverage of naturally occurring MT outputs, domains, and language pairs remains an important direction. More broadly, BAE-Series can potentially serve not only as an evaluator but also as a reference-free reward model for translation post-training, enabling future work on whether rubric-guided optimization can improve both translation correctness and finer-grained quality beyond correctness.

# 6 Conclusion

We introduced BAE, a reference-free method for evaluating translation quality beyond semantic adequacy. We first design BAE-Rubric, which organizes translation quality into three core dimensions and four dimensions beyond semantic adequacy. Building on these criteria, we construct BAE-Bench, which combines preference-graded training data with an independently human-annotated test set. Finally, BAE-Series learns rubric-guided scoring from this benchmark through SFT and GRPO. The resulting evaluators produce a quality score together with a multidimensional rationale, offering a more interpretable basis for evaluating increasingly capable machine translation systems.

### AI USE STATEMENT

AI tools were used to assist with code implementation, debugging, and surveying related literature, as well as to improve the readability and language of the manuscript. All AI-assisted content was reviewed and verified by the authors. LLMs were also used to generate synthetic data; the specific procedure is described in the paper. The authors take full responsibility for the final content of this work, including any text or artifacts produced with the assistance of generative AI.

### ETHICS STATEMENT

Our Tier-5 anchors used in this work are drawn from publicly available sources, and copyright of the original authors is respected. Tier 1–4 examples are synthetically constructed via controlled LLM-based degradation of these anchors, and the resulting BAE-BENCH test set was subsequently reviewed and annotated by human raters (Appendix D).

### REPRODUCIBILITY STATEMENT

We describe BAE-RUBRIC in Section 2.2 and Appendix C; the data-construction and post-training pipeline, including the degradation procedure, teacher models, and the SFT and GRPO objectives, in Sections 2.3–2.4 and Appendices A–B; and all evaluation prompts and baseline model versions in Appendices F and E.1. Upon acceptance, we will release the code, the BAE-BENCH, the trained BAE-SERIES checkpoints, and the training and evaluation scripts used to produce the results in this paper.

## REFERENCES

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. Can LLMs really learn to translate a low-resource language from one grammar book? In *International Conference on Learning Representations*, 2025.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge, 2001.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 1066–1083, 2023.

Yves Ferstler, Adam Podoxin, Ty Brassington, Gaëlle Laperrière, Roman Grundkiewicz, Marie-Jean Meurs, Maite Taboada, and Marzena Karpinska. Ai translation of literary texts is "fine", but readers still prefer human translations. 2026. URL https://arxiv.org/abs/2606.26040.

Mara Finkelstein, Daniel Deutsch, Parker Riley, Juraj Juraska, Geza Kovacs, and Markus Freitag. From jack of all trades to master of one: Specializing LLM-based autoraters to a test set. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17221–17238, 2025.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. You cannot feed two birds with one score: The accuracy–naturalness tradeoff in translation. *arXiv preprint arXiv:2503.24013*, 2025.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81, Miami,

Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.2. URL https://aclanthology.org/2024.wmt-1.2/.

Wenyang Gao, Yinghao Yang, Xi Jin, Jing Li, and Yue Zhang. FiRE: Fine-grained ranking evaluation for machine translation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

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

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, 2024.

Ahrii Kim. RUBRIC-MQM: Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Industry Track*, pp. 147–165, 2025.

Junghwan Kim, Kieun Park, Sohee Park, Hyunggug Kim, and Bongwon Suh. MAS-LitEval: Multi-agent system for literary translation quality assessment. *arXiv preprint arXiv:2506.14199*, 2025.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775, 2023.

Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. Mufu: Multilingual fused learning for low-resource translation with LLM. In *International Conference on Learning Representations*, 2025.

Yikang Liu, Wanyang Zhang, Yiming Wang, Jialong Tang, Pei Zhang, Baosong Yang, Fei Huang, Rui Wang, and Hai Hu. Translationese-index: Using likelihood ratios for graded and generalizable measurement of translationese. In *Proceedings of EMNLP*, pp. 12521–12538, 2025.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, (12):455–463, 2014. doi: 10.5565/rev/tradumatica.77.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. *arXiv preprint arXiv:2405.16969*, 2024.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Yichen Lu, Wei Dai, Jiaen Liu, Ching Wing Kwok, Zongheng Wu, Xudong Xiao, Ao Sun, Sheng Fu, Jianyuan Zhan, Yian Wang, Takatomo Saito, and Sicheng Lai. ViDove: A translation agent system with multimodal context and memory-augmented reasoning. In *Proceedings of EMNLP: System Demonstrations*, 2025.

Monishwaran Maheswaran, Marco Carini, Christian Federmann, and Tony Diaz. TASER: Translation assessment via systematic evaluation and reasoning. In *Proceedings of the Tenth Conference on Machine Translation*, pp. 1004–1010, 2025. doi: 10.18653/v1/2025.wmt-1.76.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. MT-Ranker: Reference-free machine translation evaluation by inter-system ranking. In *International Conference on Learning Representations*, 2024.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, 2026. System card: https://deploymentsafety.openai.com/gpt-6-astra. Accessed 2026-09-26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, 2002.

Dmitry Popov, Vladislav Negodin, Ekaterina Enikeeva, Iana Matrosova, Nikolay Karpachev, and Max Ryabinin. Refined assessment for translation evaluation: Rethinking machine translation evaluation in the era of human-level systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22079–22095, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. Accessed 2026-09-26.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, 2020.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 841–848, Singapore, 2023. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of ACL*, 2020.

Sheikh Shafayat, Dongkeun Yoon, Woori Jang, Jiwoo Choi, Alice Oh, and Seohyon Jung. A 2-step framework for automated literary translation evaluation: Its promises and pitfalls. *arXiv preprint arXiv:2412.01340*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Richard Sproat, Tianyu Zhao, and Llion Jones. TransEvalnia: Reasoning-based evaluation and ranking of translations. *arXiv preprint arXiv:2507.12724*, 2025.

Shaomu Tan and Christof Monz. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. In *Proceedings of EMNLP*, pp. 4370–4387, 2025.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Learning Representations*, 2024.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. DelTA: An online document-level translation agent based on multi-level memory. In *International Conference on Learning Representations*, 2025.

Chenxi Whitehouse, Sebastian Ruder, Tony Zhiyang Lin, Oksana Kurylo, Haruka Takagi, Janice Lam, Nicolò Busetto, and Denise Diaz. MENLO: From preferences to proficiency – evaluating and modeling native-like quality across 47 languages. In *International Conference on Learning Representations*, 2026.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922, 2025. doi: 10.1162/tacl.a.25.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations*, 2024a.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of ICML*, 2024b.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *International Conference on Learning Representations*, 2025.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of EMNLP*, pp. 29099–29121, 2025.

Xiying Zhao, Zhoufutu Wen, Zhixuan Chen, Jingzhe Ding, Jianpeng Jiao, Shuai Li, Xi Li, Danni Liang, Shengda Long, Qianqian Liu, Xianbo Wu, Hongwan Gao, Xiang Gao, Liang Hu, Jiashuo Liu, Mengyun Liu, Weiran Shi, Chenghao Yang, Qianyu Yang, Xuanliang Zhang, Ge Zhang, and Wenhao Huang. DiscoX: Benchmarking discourse-level translation in expert domains. In *International Conference on Learning Representations*, 2026.

# A Training Details

## A.1 Pipeline

**Algorithm 1** BAE Pipeline

**Require:** high-quality source–translation pairs $\mathcal{D}_5$, evaluation rubric $\mathcal{R}$, base evaluator $\mathcal{M}_\theta$, transformation model $\mathcal{G}_d$, annotation teacher $\mathcal{G}_a$, number of GRPO steps $T$
**Ensure:** Trained evaluator $\mathcal{M}_{\theta^*}$

```
Initialize: D_SFT ← ∅, D_RL ← ∅
for each quality anchor (x, y^(5)) ∈ D_5 do
    {y^(1), ..., y^(4)} ← G_d(x, y^(5), R)                ▷ Controlled quality transformation
    for q = 1, ..., 5 do
        z̃^(q) ← G_a(x, y^(q), q, R)                       ▷ Label-conditioned rationale
        D_SFT ← D_SFT ∪ {(x, y^(q), q, z̃^(q))}
        D_RL ← D_RL ∪ {(x, y^(q), q)}
θ_SFT ← SFT(M_θ, D_SFT)                                   ▷ Learn structured evaluation
θ ← θ_SFT
for t = 1, ..., T do                                      ▷ T GRPO optimization steps
    Sample a minibatch B_t ⊂ D_RL
    for each (x, y, q*) ∈ B_t do
        {τ_j}_{j=1}^G ← M_θ(x, y, R), where τ_j = (q̂_j, z_j)
        for j = 1, ..., G do
            R_j ← R_score(q̂_j, q*)                        ▷ Outcome-level reward
    θ ← GRPO(θ, {(τ_j, R_j)}_{(x,y,q*)∈B_t})
θ* ← θ
return M_θ*
```

## A.2 Supervised Fine-Tuning and Reinforcement Learning

Both stages use the same pointwise input prompt (Prompt 4).

Table 5: Training hyperparameters of BAE-9B. A dash indicates a hyperparameter not applicable to that stage.

| Hyperparameter | SFT | GRPO |
|---|---|---|
| Initialization | Qwen3.5-9B | SFT checkpoint (same LoRA adapter) |
| Precision | bfloat16 | bfloat16 |
| LoRA rank / $\alpha$ / dropout | 16 / 32 / 0.05 | 16 / 32 / 0.05 (continued) |
| Learning rate | $1 \times 10^{-4}$ | $3 \times 10^{-6}$ |
| LR schedule / warmup ratio | cosine / $1/30$ | linear / 0.1 |
| Optimizer | AdamW | AdamW-8bit |
| Weight decay | 0 | 0 |
| Training length | 2 epochs | 200 steps |
| Max sequence length | 2,048 | 2,048 |
| Microbatch size / gradient accumulation | 1 / 8 | 2 / 2 |
| Effective batch size | 8 | 4 |
| Score-token loss weight | 8 | — |
| Rollouts per input $G$ | — | 2 |
| KL coefficient $\beta$ | — | 0.02 |

## A.3 Hardware and Runtime

BAE-9B is trained on a single NVIDIA RTX 4090 (24 GB), and BAE-4B on a single NVIDIA L40S (48 GB). Table 6 reports wall-clock time and GPU-hours separately for SFT and GRPO.

Table 6: Wall-clock training cost. Times are hours:minutes:seconds, followed by GPU-hours in parentheses.

| Model | GPU | SFT | GRPO | Total |
|---|---|---|---|---|
| BAE-4B | L40S | 1:22:48 (1.38) | 1:40:35 (1.68) | 3:03:23 (3.06) |
| BAE-9B | RTX 4090 | 1:48:39 (1.81) | 1:54:23 (1.91) | 3:43:02 (3.72) |

# B Data Construction Details

## B.1 Controlled modification Prompt

Prompt 1 is the original Chinese prompt used to query the transformation model; we translate it into English as Prompt 2 for inclusion here. The placeholders {`metrics_guide`}, {`input_text`}, and {`output_text`} are filled with the full rubric, the English source, and the verified Tier-5 Chinese translation, respectively. Examples illucontrolled degradationstrating the full rubric are given in Appendix C.1. The double braces in the JSON example are the template's escaped literal braces.

**Prompt 1: Controlled degradation from a Tier-5 anchor**

你是一个专业的文学翻译与机器翻译质量评估专家。请严格基于给定的英文原文和标准5 分（金标准参考译文），根据Beyond-Accuracy 的完整7 大评估维度（包含S1/S2/S3/SA1/SA2/SA3/SA4 共22 个子维度）进行受控降级改写。

【项目完整评估标准体系**(metrics.md)**】：
<metrics>
{metrics_guide}
</metrics>

【待处理数据】：
【英文原文】：{input_text}
【5分参考译文(Ground Truth)】：{output_text}

【任务指令】：
请基于上述完整评估标准体系，以5 分参考译文为基准（结合英文原文锚定），受控改写/降级生成4分、3分、2分、1分四个不同质量梯度的中文译文。改写必须精准体现各分值梯度的质量差异，严禁随意乱生成无意义文本或过度偏离：

**1. 4**分**(Good /** 仅存在轻微可见缺陷**).** 硬性要求：必须完整保留S1 语义准确性、S2 语言质量和S3 功能对等。不得出现任何误译、漏译、增译、关系错误或术语误用；不得出现语法、用词、语序或可读性退化；不得削弱原文意图、交际功能、语气、情态或语用力量。降级方向：仅允许削弱SA 维度。措辞更平淡泛化，选择更显性、更普通的句式连接；减少地道搭配和自然韵律；削弱语域与体裁契合度；略微弱化情绪色彩和正式度控制。整体读起来像一个称职但明显不够精炼的译者初稿——正确而自然，但不够凝练、不够有神韵、不够锐利。边界：与5 分GT 的差距应清晰可辨，但译文本身必须完全正确、通顺、可理解。严禁引入任何S1/S2/S3 层面的退化。

**2. 3**分**(Usable /** 可用但明显有缺陷**).** 硬性要求：保留主要主题和大部分核心命题，但译文需呈现跨S1/S2/S3 与SA 各维度的中等质量水平。降级方向：引入若干可恢复的指标缺陷，包括：次要细节、修饰语或修辞细微之处的轻度至中度遗漏；关系、施事、时序关联或文学术语的偶尔不准确渲染；别扭但仍可读的语法、分句衔接、语序或搭配；交际力量、语气、语域、意象感和韵律的明显削弱。边界：整体可理解且可辨识为基于原文的翻译，但不可出版或达到精细翻译标准。必须明显劣于4 分，明显优于2 分。严禁出现灾难性含义丢失、丢弃大部分原文、产生乱码或不可读文本。

**3. 2**分**(Poor /** 质量差但仍部分可恢复**).** 硬性要求：仍可辨识为原文的翻译，但必须包含多处明显的指标失败，绝不可作为合格译文。降级方向：引入跨S1/S2/S3 的多处中等问题，包括：限定语、修饰语或次要命题的部分遗漏；非核心细节的偶尔误译；生硬的语序、僵硬的直译措辞或不自然搭配；语气、语域、修辞力量或语篇契合度的削弱。边界：保留足够的原文含义，使整体主题和若干核心命题仍可恢复。必须劣于3 分，完全不可接受为合格译文。严禁使译文与原文无关、丢弃大部分原文内容或产生乱码。

**4. 1**分**(Severe Failure /** 严重翻译失败**).** 硬性要求：看起来是一个严重失败的翻译尝试(failed translation attempt)，绝非与原文无关的随机文本或纯机械乱码。降级方向：引入跨S1/S2/S3 的

实质性失败，包括：重要命题的遗漏；关系、施事、修饰语或时序/逻辑关联的错误；硬套直译、破碎搭配和不自然的目标语结构；语气、交际功能和语境契合度的严重丧失。边界：仅保留与原文主题的粗略联系（确保可辨识为同一条数据的失败翻译），必须明显劣于2 分，完全不可用作参考译文。严禁返回纯碎片、空白文本或纯乱码。

【通用禁止规则】：
严禁生成完全脱离原文内容的随机文本、纯乱码或截断残句；
严禁在译文前后添加任何标题、作者名或解释说明文字；
严禁改变翻译方向（必须输出中文）。

【输出格式要求】：
请仅返回一个标准的JSON 对象，格式必须完全符合如下结构，禁止包含任何Markdown 格式以外的解释说明。注意：译文内容中若包含对话或引用，请务必使用中文全角双引号“ ”或转义字符\"，绝不能出现未转义的半角英文双引号”，以免破坏JSON 格式！

```
{{
 "4":  "4分译文...",
 "3":  "3分译文...",
 "2":  "2分译文...",
 "1":  "1分译文..."
}}
```

**Prompt 2: Controlled degradation from a Tier-5 anchor (English translation)**

You are a professional literary translator and machine-translation quality evaluation expert. Based strictly on the given English source and the standard 5-point (gold-standard reference) translation, perform a controlled degradation rewrite according to the complete Beyond-Accuracy framework of 7 evaluation dimensions (S1/S2/S3/SA1/SA2/SA3/SA4, comprising 22 sub-dimensions in total).

**[Full Project Evaluation Standard (metrics.md)]:**
<metrics>
{metrics_guide}
</metrics>

**[Data to Process]:**
[English Source]: {input_text}
[5-Point Reference Translation (Ground Truth)]: {output_text}

**[Task Instruction]:**
Based on the complete evaluation standard above, using the 5-point reference translation as the baseline (anchored jointly with the English source), perform a controlled rewrite/degradation to generate four Chinese translations at tiers 4, 3, 2, and 1. The rewrites must precisely reflect the quality differences at each level; randomly generating meaningless text or deviating excessively is strictly forbidden:

**1. Score 4 (Good / only minor, visible flaws).** Hard requirement: S1 semantic accuracy, S2 language quality, and S3 functional equivalence must be fully preserved. No mistranslation, omission, addition, relation error, or terminology misuse may be introduced; no degradation in grammar, word choice, word order, or readability is allowed; the original's intent, communicative function, tone, modality, or pragmatic force must not be weakened. Direction of degradation: only the SA dimensions may be weakened. Wording becomes flatter and more generic, favoring more explicit, more ordinary sentence connectives; idiomatic collocations and natural rhythm are reduced; register and genre fit are weakened; emotional coloring and control of formality are slightly weakened. Overall it should read like a competent but noticeably unpolished translator's first draft — correct and natural, but lacking concision, flair, and sharpness. Boundary: the gap from the 5-point GT should be clearly discernible, but the translation itself must remain fully correct, fluent, and comprehensible. Introducing any degradation at the S1/S2/S3 level is strictly forbidden.

**2. Score 3 (Usable / usable but with noticeable flaws).** Hard requirement: the main topic and most core propositions must be preserved, but the translation should show medium-tiers across the S1/S2/S3 and SA dimensions. Direction of degradation: introduce several recoverable metric defects, including: light-to-moderate omission of minor details, modifiers, or rhetorical nuances; occasional inaccurate rendering of relations, agency, temporal links, or literary terminology; grammar, clause connection, word order, or collocation that is awkward but still readable; noticeable weakening of communicative force, tone, register, imagery, and rhythm. Boundary: the whole should be comprehensible and recognizable as a translation of the source, but not publishable or up to refined-translation standards. It

must be clearly worse than score 4 and clearly better than score 2. Catastrophic loss of meaning, discarding most of the source, garbled text, or unreadable text is strictly forbidden.

**3. Score 2 (Poor / poor quality but still partly recoverable).** Hard requirement: still recognizable as a translation of the source, but must contain multiple clear metric failures and can in no way count as an acceptable translation. Direction of degradation: introduce multiple moderate problems across S1/S2/S3, including: partial omission of qualifiers, modifiers, or secondary propositions; occasional mistranslation of non-core details; stiff word order, rigid literal wording, or unnatural collocations; weakened tone, register, rhetorical force, or discourse fit. Boundary: enough of the source meaning must remain that the overall topic and some core propositions are still recoverable. It must be worse than score 3 and completely unacceptable as a qualified translation. Making the translation unrelated to the source, discarding most of the source content, or producing garbled text is strictly forbidden.

**4. Score 1 (Severe Failure / severe translation failure).** Hard requirement: it should look like a severely failed translation attempt, never random text unrelated to the source or pure mechanical gibberish. Direction of degradation: introduce substantive failures across S1/S2/S3, including: omission of important propositions; errors in relations, agency, modifiers, or temporal/logical links; forced literal translation, broken collocations, and unnatural target-language structure; severe loss of tone, communicative function, and contextual fit. Boundary: only a rough connection to the source topic needs to remain (so it is still recognizable as a failed translation of the same data item); it must be clearly worse than score 2 and completely unusable as a reference translation. Returning pure fragments, blank text, or pure gibberish is strictly forbidden.

**[General Prohibitions]:**
Generating random text, pure gibberish, or truncated fragments completely detached from the source content is strictly forbidden;
Adding any title, author name, or explanatory text before or after the translation is strictly forbidden;
Changing the translation direction (Chinese output is mandatory) is strictly forbidden.

**[Output Format Requirement]:**
Return only a single standard JSON object, whose format must exactly match the structure below; no explanatory text outside standard JSON is allowed. Note: if the translation content contains dialogue or quotations, be sure to use full-width Chinese double quotation marks “ ” or the escape sequence \", and never leave an unescaped half-width English double quote " in the string, as it would break the JSON format!

```
{{
 "4":  "tier-4 translation...",
 "3":  "tier-3 translation...",
 "2":  "tier-2 translation...",
 "1":  "tier-1 translation..."
}}
```

### B.2 Label-Conditioned Rationale Generation Prompt

The placeholder {source} is the English source, and {candidate_1}–{candidate_5} are the five tiers' Chinese translations. The teacher returns the rationale split by dimension so that humans can first check each dimension's justification for plausibility and reliability; before training, the seven dimension sentences are merged into the single rationale string shown in Appendix B.4.

**Prompt 3: Label-conditioned rationale annotation**

**[System]**

You are annotating a translation-quality dataset for training an MT evaluator. For each source segment you are given five candidate Chinese translations whose true quality bands (1 = worst, 5 = best) are already established by human annotation. Your job is NOT to score them — the scores are given. Your job is to write the justification: for each candidate, one English sentence per dimension explaining what that dimension looks like in that specific candidate.

Dimensions: **S1** semantic accuracy and literal error control; **S2** target-language grammar, fluency, and readability; **S3** functional equivalence, intent, and pragmatic clarity; **SA1** native-like expression and collocation; **SA2** stylistic refinement and expressive quality; **SA3** register, genre, and discourse-context fit; **SA4** tone, formality, and attitude consistency.

Rules:

— Write in English. Quote Chinese or English fragments from the texts when useful.
— Each sentence must be concrete and specific to that candidate. Name the actual phrase that succeeds or fails. Never write generic filler.
— Do NOT state any numeric score anywhere. No "scored N points", no digits used as ratings. The band is already known; you are explaining it, not restating it.
— The five candidates of one segment differ in quality. Make the contrast visible: a band-5 justification should read clearly better than a band-4 one for the same source, and a band-1 one clearly worse than band-2.
— Return strict JSON only, no markdown fences.

**[User]**

Write the per-dimension justifications for the following segment.

Source (English): {source}

Candidate band 1: {candidate_1}
Candidate band 2: {candidate_2}
Candidate band 3: {candidate_3}
Candidate band 4: {candidate_4}
Candidate band 5: {candidate_5}

Return a JSON object of the form: {{"candidates": [{{"band": <1-5>, "S1": "...", "S2": "...", "S3": "...", "SA1": "...", "SA2": "...", "SA3": "...", "SA4": "..."}}]}}. Include every candidate you were given, with the same bands.

### B.3 Data Splits

The resulting training split contains 6,960 examples. The frozen test set is disjoint from it: 505 candidates drawn from 101 source groups, labeled through the annotation, review, and adjudication process in Appendix D.2. The annotator and reviewer did not see the constructed tier labels or the intended ordering.

### B.4 Training Example with Rationale

Table 7 shows one five-tier training group on BAE-Bench. The complete label-conditioned teacher rationales used as SFT targets are reproduced below.

Table 7: Five-tier training example from BAE-Bench.

**Source:** "The only completely stationary object in the room was an enormous couch on which two young women were buoyed up as though upon an anchored balloon. They were both in white, and their dresses were rippling and fluttering as if they had just been blown back in after a short flight around the house."

| Training label | Chinese translation |
|---|---|
| 1 | 屋里完全静止一个大沙发，上面两个年轻女人，像气球挂着。她们白衣服，衣服飘，刚短飞一圈后被吹回。 |
| 2 | 在房间里，完全静止的唯一事物是一张大沙发，两个年轻女人在它上面，好像飘在一个固定气球上。她们都穿着白色，她们的衣服在波动和飘扬，好像她们做了一个在房子周围的短飞行，刚被吹进来。 |
| 3 | 屋子里面，一个很大的沙发是唯一没有在动的东西，两个年轻女孩浮在上面，就像它被一只气球固定住了似的。这两个人都穿着白颜色的衣服，裙子在摆动，仿佛她们刚在房子周围飞完了很短的那么一圈，然后被风给吹了进来。 |
| 4 | 房间里唯一完全不动的东西，是一张巨大的沙发。两个年轻女子飘在上面，就好像被一只锚定在地上的气球托着。她们都穿着白色，裙子在风中飘动着，好像她们刚刚在房间里飞了很短的一圈，才被吹了回来。 |
| 5 | 室内唯一完全静止的是一张巨大沙发，两个年轻女子飘浮其上，仿佛挂在下锚的热气球上。她们都穿白衣，洋装在风中飘动，好似她们方才短暂在屋里飞了一圈，才刚被吹回来。 |

**Training label 1: complete teacher rationale**

The translation preserves the basic fragments of the room, the large sofa, two young women in white, fluttering clothes, and being blown back, but omits key details such as ”only,” ”enormous,” ”anchored balloon,” and ”around the house,” and mistranslates ”buoyed up” as ”气球挂着” (”hung by a balloon”).
The translation has serious grammatical problems, such as ”屋里完全静止一个大沙发” (”in the room completely still a large sofa”), ”上面两个年轻女人” (”on it two young women”), and ”刚短飞一圈后被吹回” (”after a short flight around was blown back”), with fragmentary syntax that does not match Chinese expression.
Readers can only roughly guess that there is a sofa and two women in the room; the original’s continuous imagery of stillness, floating, and fluttering clothes is hard to understand clearly.
The expression is highly unnatural; collocations such as ”完全静止一个大沙发” (”completely still a large sofa”) and ”刚短飞一圈” (”just made a short flight around”) clearly do not read like native expression, and the rhythm is broken.
The original picture is gorgeous and fluid; the translation leaves only fragmentary literal renderings, with almost no effective visual quality or stylistic polish.
The translation does not fit the register of literary narration; it reads like compressed word blocks, and its generic fit is very poor.
The original’s light, dreamy atmosphere is basically lost; the translation is mechanical and crude, failing to preserve the emotional coloring.

**Training label 2: complete teacher rationale**

The translation retains the main information of the only still thing in the room, the large sofa, the two young women in white, and the fluttering clothes, but ”anchored balloon” is mistranslated as ”固定气球” (”a fixed balloon”), ”buoyed up” is not fully rendered, and the relation of ”blown back in after a short flight around the house” is expressed clumsily and partly distorted.
The grammar is basically readable, but ”两个年轻女人在它上面” (”two young women on it”), ”衣服在波动和飘扬” (”clothes waving and fluttering”), and ”做了一个短飞行” (”made a short flight”) are clearly Europeanized in wording and structure.
Readers can largely understand the scene, but several metaphors and action relations are vague, especially the relation between the flight and being blown back in the last sentence, which is not clear enough.
The native feel is weak; ”完全静止的唯一事物” (”the only completely still thing”), ”固定气球” (”a fixed balloon”), and ”短飞行” (”a short flight”) are stiff collocations, and the sentence flow is insufficient.
The original has strong pictorial quality; the translation keeps some visual elements but expresses them clumsily, failing to form an elegant, vivid image.
The translation can still count as novelistic description, but it is overly literal, and both the literary register and the Fitzgerald-like lightness are clearly insufficient.
The original’s light, dreamy, slightly exaggerated mood is partially visible but weakened by mechanical expression, and the emotional coloring is not quite on target.

**Training label 3: complete teacher rationale**

The translation basically preserves the still sofa in the room, the two women, the white clothes, and the image of flying back, but ”as though upon an anchored balloon” is misinterpreted as the sofa being fixed by a balloon, with an obvious deviation in the relation.
The grammar is largely workable, but words such as ”白颜色的衣服” (”white-colored clothes”) and ”很短的那么一圈” (”such a short little circle”) are verbose and unnatural.
The overall scene can be understood, but the balloon metaphor misleads readers, and part of the intended meaning is unclear.
The expression is colloquial and heavy; ”浮在上面” (”floating on top”) and ”白颜色的衣服” (”white-colored clothes”) are not literary-native enough.
Some images of flying and fluttering hems are retained, but the metaphor is mishandled and the style is not refined enough.
Within the register of novelistic description, the translation is plain and clumsy, and the generic fit is average.
The light, drifting emotion is partly preserved but weakened by the mistranslated relation and colloquial expression.

**Training label 4: complete teacher rationale**

The translation basically and accurately conveys the still sofa in the room, the two women in white, being held up by a balloon, and the fluttering clothes, but "around the house" rendered as "在房间里" ("in the room") narrows the scope.
The syntax is smooth, the word choice is basically natural, "飘在上面" ("floating on top") is slightly exaggerated but consistent with the original meaning, and overall readability is good.
Readers can clearly understand the picture and the metaphorical relation, the logic between the sentences is clear, and there is no obvious ambiguity.
"被一只锚定在地上的气球托着" ("held up by a balloon anchored to the ground") and "裙子在风中飘动着" ("the skirts fluttering in the wind") quite match Chinese expression, with only a few short phrases slightly stiff.
The translation effectively preserves the images of floating, white skirts, and being blown back after a brief flight, with strong literariness, though "房间里飞了一圈" ("flew a circle in the room") slightly weakens the original's sense of space.
Overall it suits the register of novelistic description, and the imagery and sentence patterns accord with literary narration.
The original's light, strange atmosphere is basically preserved, the emotional coloring is appropriate, and there is no excess of formality or deviation in attitude.

**Training label 5: complete teacher rationale**

The translation basically preserves the sofa, the two women, the white clothes, the fluttering skirts, and the flying-back metaphor, but "around the house" rendered as "在屋里" ("in the room") has a deviation in the range of location.
The grammar is correct, the sentences are coherent, and the word order is natural.
Readers can clearly understand the picture of the still sofa in the room and the women floating up like balloons, and the core intention is clear.
"飘浮其上" ("floating upon it") and "方才短暂……才刚被吹回来" ("just briefly ... just been blown back") have a literary expressiveness, but "挂在下锚的热气球上" ("hung on an anchored hot-air balloon") is slightly stiff.
The translation nicely presents the light, floating, flying imagery, with a strong pictorial quality.
Overall it fits the register of novelistic description, and the slightly classical "其上" ("upon it") and "方才" ("just now") also serve literary narration.
The original's dreamy, airy atmosphere is basically preserved, and the formality suits literary narration.

# C BAE-Rubric Definitions and Illustrative Cases

## C.1 Illustrative Chinese Cases and Human Check Focus

The cases below illustrate the full rubric for English-to-Chinese translation. A *wrong* candidate illustrates a defect; a *better/baseline* pair illustrates a relative difference. For source-free examples, only target-language quality can be judged. The human check focus is the criterion annotators inspect, not an additional numerical label.

**S1: Accuracy.**

- **`S-1.1` Mistranslation.** Source: "He demanded an apology." Wrong: "他请求道歉。" Check whether the demand has been weakened into a request (是否把原文意思翻错).
- **`S-1.2` Omission.** Source: "three large wooden crates." Wrong: "三个箱子." Check whether *large* and *wooden* are missing (是否漏掉原文关键信息).
- **`S-1.3` Addition.** Source: "She smiled." Wrong: "她高兴地笑了。" Check whether the emotion was added without support from the source (是否无依据增译).
- **`S-1.4` Misrepresented Relations.** Source: "John hit Peter." Wrong: "彼得打了约翰。" Check participant roles and other logical or temporal relations (人物关系、时序、逻辑是否错).
- **`S-1.5` Terminology Misuse.** Source: "machine learning." Wrong: "机械学习." Check whether the key domain term is translated correctly (术语或领域概念是否错).

**S2: Language Quality.**

- **S-2.1 Basic Grammar Correctness.** Wrong target-language text: “他明天去会。” Check whether the sentence contains a clear grammatical error (语法是否明显错误).
- **S-2.2 Lexical Appropriateness.** Wrong target-language text: “这部电影非常吃。” Check for inappropriate or misused words (词汇选择是否不自然或误用).
- **S-2.3 Natural Word Order.** Wrong target-language text: “我昨天看了电影在家。” Check whether the target-language ordering is natural (语序是否自然).
- **S-2.4 Coherence and Readability.** Wrong target-language text: “这个方案的优点它没有明显的缺点但我们也不确定它是否适合。” Check whether the sentence is hard to read or confusing (是否难读、别扭、混乱).

**S3: Functional Equivalence.**

- **S-3.1 Clarity of Intended Meaning.** Ambiguous target-language text: “她告诉她她会来。” Check whether the intended participants can be identified clearly (是否能清楚理解intended meaning).
- **S-3.2 Communicative Function Preservation.** Source: “Do not enter.” Wrong: “不要进入吗？” Check whether a warning or command was turned into a question (疑问、警告、请求等功能是否保留).
- **S-3.3 Pragmatic Interpretability.** Source: “You might want to reconsider.” Wrong: “你最好重新考虑！” Check whether the tentative modality was strengthened (语气、情态、意图是否偏差).

**SA1: Native and Natural Expression.**

- **SA-1.1 Native-like Expression.** Source: “The plan fell apart.” Better: “计划泡汤了。” Baseline: “计划散开了。” Check whether the phrase reads as native expression (是否像母语表达).
- **SA-1.2 Natural Collocations.** Source: “strong evidence.” Better: “有力的证据.” Baseline: “强壮的证据.” Check the target-language collocation (搭配是否地道).
- **SA-1.3 Smooth Rhythm and Flow.** Source: “He paused for a moment.” Better: “他停顿了一下。” Baseline: “他暂停了一刻。” Check rhythm and fluency (节奏和流畅度是否自然).

**SA2: Stylistic Refinement.**

- **SA-2.1 Vivid Imagery.** Source: “The city grew quiet.” Better: “城市渐渐安静下来。” Baseline: “城市变得安静。” Check whether the target preserves the scene without adding new meaning (是否更有画面感但不增义).
- **SA-2.2 Stylistic Refinement.** Source: “He spoke softly.” Better: “他轻声说道。” Baseline: “他轻轻地说话。” Check whether refinement preserves the source meaning (文风是否更refined).

**SA3: Register and Genre Fit.**

- **SA-3.1 Discourse Register.** Literary source: “What happens when the devil walks among us?” Baseline: “当魔鬼行走在人间会发生什么？” A more literary rendering is “若魔鬼混迹人间，会发生什么？” Check register fit without introducing an unsupported participant or scene (是否符合语域).
- **SA-3.2 Domain and Genre Suitability.** News source: “Officials declined to comment.” Better: “官员拒绝置评。” Baseline: “官员不想评论。” Check whether wording fits news reporting (是否符合体裁、领域).

**SA4: Tone, Formality, and Attitude.**

- **SA-4.1 Emotional Tone.** Source: “I really appreciate your help.” Better: “非常感谢你的帮助。” Baseline: “我对你的帮助表示感谢。” Check whether the emotional tone comes through (情绪色彩是否到位).
- **SA-4.2 Formality.** Formal source: “We apologize for the inconvenience.” Better: “对造成的不便我们深表歉意。” Baseline: “抱歉给你带来麻烦。” Check the expected politeness level (正式度是否匹配).

- **SA-4.3 Attitude and Intent.** Source: "You might want to be careful." Better: "你可能要小心一点。" Baseline: "你最好小心一点！" Check whether caution has been turned into a stronger command (态度、讽刺、谨慎等是否保持).

# D Human Annotation

## D.1 Annotation Guidelines

Annotation is conducted directly under BAE-Rubric (Appendix C): annotators judge each candidate against the aspects defined there rather than a separate annotation scheme, and record a single overall quality grade. Annotators do not produce per-aspect scores; the rubric serves as shared vocabulary for the judgment and its justification.

## D.2 Annotator Qualifications

We followed the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001),[1] a guideline used to describe the achievements of learners of foreign languages across Europe and in other countries, and listed the proficiency of the annotator and reviewer in Table 8. The six levels in the CEFR are described as follows:

- **A1 (Breakthrough).** Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
- **A2 (Waystage).** Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
- **B1 (Threshold).** Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes ambitions and briefly give reasons and explanations for opinions and plans.
- **B2 (Vantage).** Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
- **C1 (Advanced).** Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
- **C2 (Mastery).** Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Table 8: English and Chinese proficiency of the annotator, reviewer, and rubric feedback providers based on CEFR levels.

| Language | Annotator | Reviewer | Rubric feedback providers |
|---|---|---|---|
| English | B2 | C1 | C1 |
| Chinese | C2 | C2 | C2 |

[1] Council of Europe: Common European Framework of Reference for Languages.

**Annotation, review, and adjudication.** Annotation was carried out by one annotator with B2 English proficiency and one reviewer with C1 English proficiency from a translation annotation vendor. Both were briefed on BAE-RUBRIC (Appendix C). The annotator assigned an overall quality grade to each candidate, and the reviewer checked the annotations. For particularly contentious items, they discussed the judgments and reached a final answer that both agreed on. The final labels obtained through this complete annotation, review, and adjudication process serve as the gold standard for the test set.

# E EVALUATION DETAILS

## E.1 BASELINE MODEL VERSIONS

Table 9 lists the evaluators compared in Table 3, together with their type and version or backbone, as well as the teacher model used for data construction. API-based LLM baselines are queried with high reasoning effort wherever the interface exposes such a setting; we keep this fixed across the rubric and no-rubric conditions, so that the two rows differ only in the prompt.

Table 9: Details of the models used in our experiments, including the evaluators and the teacher model used for data construction.

| Model | Type | Version / Backbone |
|---|---|---|
| *Evaluators* | | |
| COMET-QE-wmt20 | learned | `wmt20-comet-qe-da` |
| Kiwi-XXL | learned | `wmt23-cometkiwi-da-xxl` |
| xCOMET-XXL | learned | `Unbabel/xCOMET-XXL` |
| MetricX-24-XXL | learned | `metricx-24-hybrid-xxl` |
| MT-Ranker-XXL | ranking | `ibraheemmoosa/mt-ranker-xxl` |
| FiRE | LLM | `Qwen/Qwen3.5-9B` |
| GEMBA-MQM | LLM | `Qwen/Qwen3.5-9B` |
| Rubric-MQM | LLM | `Qwen/Qwen3.5-9B` |
| EAPrompt | LLM | `Qwen/Qwen3.5-9B` |
| AutoMQM | LLM | `Qwen/Qwen3.5-9B` |
| Qwen3.5-9B | LLM | `Qwen/Qwen3.5-9B` |
| DeepSeek-V4.1-Flash | LLM | `deepseek-v4.1-flash` |
| GPT-6 Astra | LLM | `gpt-6-astra` |
| *Data construction* | | |
| transformation / annotation teacher | LLM | `deepseek-v4-flash` (version 0731) |

**Protocol notes.** Locally served evaluators (Qwen3.5-9B, FiRE, GEMBA-MQM) run in bfloat16.

**Quantile mapping for MAE.** Predictions tied at the same score are assigned the median human label over their tied rank range, rather than an arbitrary tie-break order.

## E.2 TEST SET EVALUATION CASE

Tables 10 and 11 show an evaluation case on BAE-BENCH and the model predictions. The construction labels and human grades agree for all five candidates in this group. The complete BAE-9B GRPO rationales for all five candidates are reproduced below.

Table 10: BAE-BENCH evaluation case.

**Source:** It was dawn now on Long Island and we went about opening the rest of the windows downstairs, filling the house with grey-turning, gold-turning light. The shadow of a tree fell abruptly across the dew and ghostly birds began to sing among the blue leaves. There was a slow, pleasant movement in the air, scarcely a wind, promising a cool, lovely day.

| Human grade | Chinese translation |
|---|---|
| 1 | 在长岛的黎明时候，我们去打开剩下的楼下窗口。这房子充满了变成灰色变成金色的光。一个树荫很快落下在露水上，鬼鸟开始在蓝叶子唱。有慢慢愉快的空气移动，几乎不是风，它允许了凉爽和漂亮的一天。 |
| 2 | 长岛的早晨到了，我们把其他的楼下窗户打开，那从灰变成金的光进入了房子。树影突然落在了露水，幽灵一样的鸟在蓝色树叶之间开始唱歌。空气中有了慢慢舒服地移动，不太算风，像在答应一个凉爽和可爱的白天。 |
| 3 | 长岛到了早晨，我们打开楼下其他的窗，让金色的光装满屋子。树的影子忽然掠过露水，鸟儿在蓝叶间唱起了歌。空气慢慢地动着，带着舒服，几乎没什么风，今天大概是凉爽舒适的一天。 |
| 4 | 现在长岛已是早晨，我们把楼下其余的窗户全打开了，屋里充满了由灰变金的光线。树影骤然落在露水上，看不见的鸟儿在蓝色叶丛间开始歌唱。空气中有一股缓慢而舒服的流动，几乎算不上是风，预示着一个凉爽宜人的日子。 |
| 5 | 现在长岛已经是清晨，我们动手把楼下其余的窗户全打开，屋里充满似灰还金的光线。一棵树的影子忽然从朝露间出现，看不见踪影的小鸟在蓝色树叶间歌唱。空气中有股缓慢舒服的流动，几乎不成风，承诺着今天会是凉爽美丽的一天。 |

Table 11: Model scores for Table 10. Column numbers identify the human grade. Learned metrics retain their native continuous scales (higher is better); other rows report 1–5 scores.

| Evaluator | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| COMET-QE-wmt20 | -0.462 | -0.492 | -0.154 | 0.099 | -0.019 |
| Kiwi-XXL | 0.474 | 0.531 | 0.652 | 0.782 | 0.787 |
| xCOMET-XXL | 0.266 | 0.360 | 0.461 | 0.553 | 0.551 |
| Qwen3.5-9B | 2 | 3 | 4 | 5 | 4 |
| DeepSeek-V4.1-Flash | 2 | 3 | 3 | 4 | 4 |
| GPT-6 Astra | 1 | 3 | 3 | 5 | 4 |
| BAE-9B | 1 | 2 | 3 | 4 | 5 |

**Human grade 1, BAE score 1: complete generated rationale**

The translation preserves some basic elements like dawn, windows, and light, but it mistranslates or mangles key images such as 'ghostly birds' as '鬼鸟' and 'promising' as '允许了'.
The Chinese is awkward and often ungrammatical, especially '一个树荫很快落下在露水上' and '有慢慢愉快的空气移动'.
The scene's function as a gentle, awakening description is weakened because the imagery sounds literal and confusing rather than evocative.
Phrases like '变成灰色变成金色的光' and '蓝叶子唱' are unnatural collocations in Chinese.
The prose lacks refinement and turns the delicate visual progression into clumsy word-for-word fragments.
It does not fit the literary descriptive genre because the syntax and diction feel like a rough draft rather than polished narration.
The tranquil, hopeful tone is disrupted by awkward wording such as '鬼鸟' and '允许了凉爽和漂亮的一天'.

**Human grade 2, BAE score 2: complete generated rationale**

The translation conveys the broad events, but '那从灰变成金的光进入了房子' is literal and awkward, and '像在答应一个凉爽和可爱的白天' mistranslates 'promising' as if the air were answering a day.

The Chinese is understandable but clumsy, with phrases like '慢慢舒服地移动' and '不太算风' sounding unnatural.
The scene is only partly clear because the imagery of light, dew, and breeze is expressed in a way that feels mechanical rather than evocative.
Collocations such as '树影突然落在了露水' and '蓝色树叶' are not idiomatic Chinese.
The prose lacks refinement and turns the delicate visual transitions into stiff explanatory wording.
It only weakly fits a literary descriptive passage because the diction is too literal and awkward for the genre.
The gentle, anticipatory tone is weakened by clumsy wording like '像在答应,' which makes the attitude sound odd rather than serene.

**Human grade 3, BAE score 3: complete generated rationale**

The translation conveys the main events, but it omits the grey-to-gold transition, simplifies the ghostly birds, and renders the promise of the day as a guess with "大概是."
The Chinese is readable, though "让金色的光装满屋子" and "空气慢慢地动着" are somewhat plain and slightly awkward.
The scene is understandable as a morning opening of windows and birdsong, but the subtle atmospheric progression is weakened.
"掠过露水" is natural, while "带着舒服" is less idiomatic than "带着舒适感" or "令人舒适."
The prose is clear but lacks the layered visual beauty of the original.
It fits a general narrative register, though it is more colloquial and less refined than the source.
The tone is pleasant, but the tentative "大概是" softens the source's assured sense of a lovely day.

**Human grade 4, BAE score 4: complete generated rationale**

The translation accurately conveys the opening of the windows, the changing light, the tree shadow, the birds, and the promise of a cool day, though "ghostly birds" rendered as "看不见的鸟儿" ("invisible birds") is slightly off.
The grammar is correct, the word order is natural, and the sentences are fluent and easy to read.
The overall meaning is clear, and readers can understand the scene and the atmosphere; the communicative function is preserved.
"由灰变金的光线" ("light changing from gray to gold") and "几乎算不上是风" ("hardly counts as wind") are natural and smooth, though a few phrases are slightly literal.
The translation has a certain visual quality and rhythm, but the refinement of "ghostly" and the delicacy of the light are not fully conveyed.
It fits the register of literary description, and the prose is fairly elegant and matches the novelistic scene.
The original's quiet, pleasant, and expectant mood is basically preserved, and the formality is appropriate.

**Human grade 5, BAE score 5: complete generated rationale**

The translation accurately conveys the opening of the windows, the gray-to-gold light, the tree shadow, the birds, and the promise of a cool day, with only minor interpretive shifts such as '看不见踪影的小鸟' for 'ghostly birds'.
The Chinese is fluent and well structured, with natural phrasing like '动手把楼下其余的窗户全打开' and '从朝露间出现'.
The passage clearly performs the same function of transforming the house and atmosphere into a bright, gentle morning scene.
Collocations such as '似灰还金的光线', '朝露间', and '凉爽美丽的一天' are idiomatic and evocative.
The prose is polished and lyrical, preserving the source's soft visual transitions and airy rhythm.
It fits the literary descriptive genre well, using elevated but accessible language appropriate to a novelistic dawn scene.
The tone remains tranquil, pleasant, and subtly anticipatory, matching the source's attitude toward the coming day.

# F PROMPT TEMPLATES

## F.1 POINTWISE EVALUATION

Prompt 4 is shared by SFT, GRPO, and pointwise evaluation with BAE-RUBRIC. Prompt 5 is the direct pointwise template. Braced fields denote input placeholders; the system and user messages are shown separately.

**Prompt 4: Pointwise evaluation with BAE-RUBRIC**

**[System]**

You are a translation quality evaluator for beyond-accuracy MT evaluation.

Given a source segment, a candidate translation, judge the candidate on these aspects:

**S1** semantic accuracy and literal error control;
**S2** target-language grammar, fluency, and readability;
**S3** functional equivalence, intent, and pragmatic clarity;
**SA1** native-like expression and collocation;
**SA2** stylistic refinement and expressive quality;
**SA3** register, genre, and discourse-context fit;
**SA4** tone, formality, and attitude consistency.

Use the context to decide the relative importance of accuracy, fluency, style, register, and tone. Discuss each aspect in the rationale, then give one overall quality band from 1 (worst) to 5 (best).

Return strict JSON only, with keys: overall, nominal_label, rationale.

**[User]**

Evaluate this candidate translation. Use the context to balance semantic correctness, naturalness, style, register, and tone.

Direction: {`direction`}
Source text: {`source`}
Candidate translation: {`candidate`}

**Prompt 5: Pointwise evaluation (short)**

**[System]**

You are a translation quality evaluator for beyond-accuracy MT evaluation.

Given a source segment, a candidate translation, judge the candidate. give one overall quality band from 1 (worst) to 5 (best).

Return strict JSON only, with keys: overall, rationale.

**[User]**

Evaluate this candidate translation.

Direction: {`direction`}
Source text: {`source`}
Candidate translation: {`candidate`}

### F.2 PAIRWISE EVALUATION

Prompts 6 and 7 compare two candidate translations of the same source, with and without the rubric, respectively.

**Prompt 6: Pairwise evaluation with BAE-RUBRIC**

**[System]**

You are a translation quality evaluator for beyond-accuracy MT evaluation.

You are given a source segment and TWO candidate translations, A and B, of that same segment. Judge them against each other on these aspects:

**S1** semantic accuracy and literal error control;
**S2** target-language grammar, fluency, and readability;
**S3** functional equivalence, intent, and pragmatic clarity;
**SA1** native-like expression and collocation;
**SA2** stylistic refinement and expressive quality;
**SA3** register, genre, and discourse-context fit;
**SA4** tone, formality, and attitude consistency.

Compare the two directly: say what each does better than the other, then score both. The two scores must differ unless the candidates are genuinely indistinguishable in quality; if one is even slightly better, it must receive the higher score.

Return strict JSON only, with keys: score_a, score_b, better, rationale. `score_a` and `score_b` are quality scores from 1 (worst) to 5 (best) and may use decimals; `better` is "A" or "B".

---

**[User]**

Compare these two candidate translations of the same source.

Source text: {source}

Translation A: {cand_a}

Translation B: {cand_b}

**Prompt 7: Pairwise evaluation (short)**

**[System]**

You are a translation quality evaluator for beyond-accuracy MT evaluation.

You are given a source segment and TWO candidate translations, A and B, of that same segment. Judge the candidates. Compare the two directly: say what each does better than the other, then score both. The two scores must differ unless the candidates are genuinely indistinguishable in quality; if one is even slightly better, it must receive the higher score.

Return strict JSON only, with keys: score_a, score_b, better, rationale. `score_a` and `score_b` are quality scores from 1 (worst) to 5 (best) and may use decimals; `better` is "A" or "B".

---

**[User]**

Compare these two candidate translations of the same source.

Source text: {source}

Translation A: {cand_a}

Translation B: {cand_b}