Lost in Mimicry: Verified Training-Free Post-Editing for LLM Translationese

Dongwon Noh[1*], Donghyeok Koh[1*], Yeonsoo Lim[2*], Seohyeong Jeong[3], Yunsu Kim[4], Gyuwan Kim[5], Soojong Do[6], JinYeong Bak[7†], Sookyoung Eo[8†], Cheoneum Park[1†]

[1]Hanbat National University, [2]LG AI Research, [3]PwC, [4]LILT, [5]UCSB,
[6]Seoul, [7]Sungkyunkwan University, [8]Yonsei University

Abstract

Translations generated by Large Language Models (LLMs) often exhibit unnatural stylistic traces distinct from human-written translations, which we refer to as LLM translationese. We propose *MELT*, a training-free post-editing framework for mitigating LLM translationese across five target languages (Korean, Chinese, Japanese, German, and French) without weight updates. MELT defines translationese patterns based on faithfulness, voice, register, naturalness, and surface form, and combines a three-stage verification gate with selective multi-agent debate to preserve meaning and mitigate self-bias. Experiments on FLORES-200 show that verification contributes most to quality improvement and that MELT achieves the lowest calibration error with respect to Oracle preference across the five languages. Our analysis further shows that LLM translationese exhibits language-specific realizations while sharing a common mechanism of English-structure mimicry.

(a) French: grammatical anglicism	
Source (En)	Travellers are strongly advised to be aware of any risk of severe weather affecting their area...
LLM (Fr)	Les voyageurs sont fortement conseillés de se renseigner sur tout risque de conditions météorologiques sévères affectant leur région...
MELT (Fr)	Il est recommandé aux voyageurs de se tenir informés des risques de conditions météorologiques sévères dans leur région...
(b) Korean: lexical calque	
Source (En)	Several million vials of encephalitis vaccine have also been promised by the government...
LLM (Ko)	정부는 내년을 대비하기 위해 여러 백만 개의 뇌염 백신 병을 약속했습니다...
MELT (Ko)	정부는 내년을 대비해 수백만 개의 뇌염 백신을 약속했습니다...

Table 1: *LLM translationese examples.* Each row pair: a single LLM call produces text whose surface looks fluent but whose construction (red) is a direct calque of English grammar / lexicon that the target language does not natively use. MELT edits are shown in blue and result in a more human-like sentence.

1 Introduction

Large Language Models (LLMs) have been widely adopted across natural language processing (NLP) tasks, including machine translation, summarization, and multilingual text generation (Brown et al., 2020; Minaee et al., 2025; Zhu et al., 2025; Qin et al., 2025; Zhao et al., 2026). LLM-generated translations often exhibit unnatural yet fluent stylistic artifacts distinct from human translations, which we refer to as LLM translationese. Despite their fluency, LLM-generated texts often exhibit subtle but measurable stylistic artifacts that distinguish them from naturally authored human text (Li et al., 2025; McGovern et al., 2025). We refer to this unnatural yet fluent style as *translationese* (Toral, 2019; Gellerstam, 1986; Baker, 1993; Volansky et al., 2015; Vanmassenhove et al., 2019). The existence and detectability of translationese have been quantified along multiple axes: including SFT-induced translationese, multilingual lexical and structural naturalness (Guo et al., 2025), PPL-curvature based detection (Mitchell et al., 2023; Hans et al., 2024), and lexical and morphosyntactic fingerprints that persist across models in the same family (McGovern et al., 2025).

Table 1 presents examples of LLM translationese. French (Fr) example, *Les voyageurs sont fortement conseillés* ... appears fluent but closely mirrors the English passive construction "Travellers are strongly advised ...". This use of *conseiller* with the advised entity as the grammatical subject is non-idiomatic in French, revealing

[*]Equal contributions.
[†]Corresponding authors.

Proceedings of the Eleventh Conference on Machine Translation, pages 462–491
October 28-29, 2026 ©2026 Association for Computational Linguistics

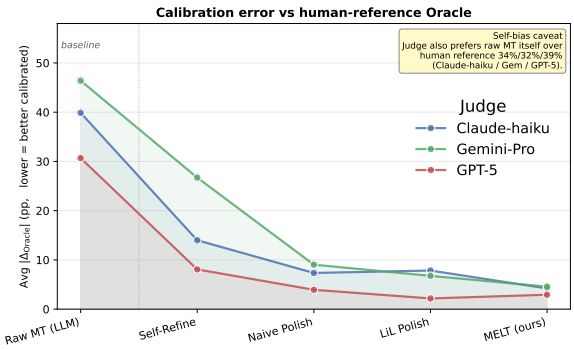

Figure 1: Calibration error of LLM-as-judge evaluation. Average $|\Delta_O|$ scores acorss five languages for each method and judge (lower = closer to human reference). Self-Refine (Madaan et al., 2023), LiL Polish (Li et al., 2025); Naive Polish is a single-prompt baseline.

an English-structured calque. Similarly, the Korean (Ko) phrase "여러 백만 개의 뇌염 백신 병", literally "several million encephalitis vaccine bottles," sounds unnatural in Korean because it mechanically maps "vials" to "병" ("bottles") and "several million" to the non-idiomatic phrase "여러 백만." Together, these examples demonstrate that surface-level fluency does not necessarily indicate target-native naturalness. Figure 1 reports the calibration of translated outputs using LLM against the oracle reference, showing that translationese introduces substantial variation in judge calibration.

Prior studies on mitigating translationese can be divided into two categories: translationese, which refers to unnatural expressions caused by source-language interference, and post-training distortion, which refers to stylistic artifacts induced by post-training processes such as RLHF or instruction tuning. For translationese, Li et al. (2025) analyze literal translation patterns in English-Chinese LLM translation and propose an SFT-based polishing model to reduce them. Guo et al. (2025) quantify English-centric lexical and syntactic transfer in multilingual LLM outputs for Chinese and French, and mitigate this effect with DPO. These studies show that LLMs can transfer English-centric expression patterns into non-English target languages, weakening native patterns of expression in the target language. For post-training distortion, Mitchell et al. (2023) and Hans et al. (2024) detect text generated by LLMs using PPL-curvature and zero-shot likelihood comparison. McGovern et al. (2025) analyze lexical and morphosyntactic fingerprints that persist across models in the same family, showing that each model family can exhibit distinguishable stylistic patterns.

However, several important gaps remain in prior studies. Existing studies have largely centered on high-resource languages, leaving typologically distant languages, such as Korean and Japanese, comparatively underexplored. In addition, most mitigation approaches rely on training-time adaptation, such as SFT or DPO, which requires additional model training and limits practical deployability. More broadly, the field lacks a unified framework for examining cross-lingual manifestations of translationese under a shared taxonomy and consistent evaluation protocol.

To address these challenges, we propose *MELT* (**M**ultilingual **E**diting for **L**LM-ese with **T**raining-free post-editing), and apply it consistently to five target languages, such as Korean, Chinese, Japanese, German, and French, with English as the source language. Figure 2 illustrates the overall architecture of MELT. To examine whether this design is effective for mitigating LLM translationese, we begin by asking the following questions: 1) How does LLM translationese manifest across different target languages? 2) Can inference-time post-editing mitigate translationese while preserving the meaning of the source text? 3) Does an LLM judge reward translationese removal as naturalness, or does it misinterpret it through self-bias?

We detect translationese patterns across five categories: faithfulness, voice, register, naturalness, and surface forms. MELT then performs conditional editing on the detected spans exhibiting translationese patterns, followed by verification of semantic preservation and naturalness through surface-level checks, bidirectional NLI, and LLM-judge-based gating. For categories with a potential risk of self-bias, MELT further applies selective debate between a universal judge and a native judge. If any gate fails, the edited output is safely rolled back to the original translation. This *detect-edit-verify* process applies the same framework to five languages in parallel without updating model weights, enabling cross-lingual comparison, inference-time mitigation, and evaluation reliability analysis within a single pipeline.

Our contributions are summarized as follows:

- We define and analyze a multilingual taxonomy of LLM translationese across five languages using five categories.
- We propose MELT, a plug-and-play post-editing framework, and show that training-

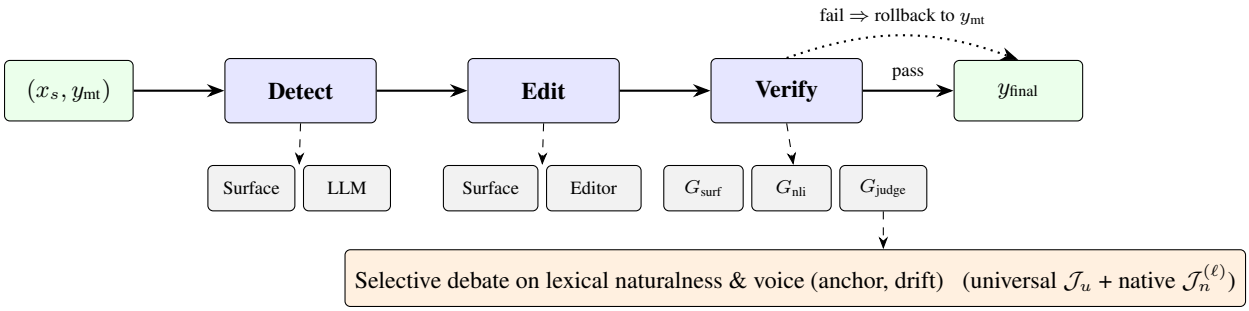

Figure 2: MELT architecture. Given an input pair (x_s, y_{mt}), the final output y_{final} is produced through detection, editing, verification, and selective debate restricted to self-bias-prone categories.

free editing can mitigate translationese while preserving meaning.

- MELT achieves the best calibration with human reference (Oracle) in terms of prefer-edit rate across five target languages, and better preserves stylometric distribution in recovery scenarios.

2 The MELT Model

We propose *MELT*, a translationese post-editing model that operates at inference time without updating model weights. MELT (i) *detects* translationese accumulated in the initial LLM translation, (ii) performs span-conditioned *editing* based on the translationese taxonomy (Section D), and (iii) *verifies* meaning preservation and naturalness through multi-stage verification gates. For categories with a high risk of self-bias, MELT selectively invokes a multi-agent debate between a universal judge and a native judge to obtain judgments that are robust to self-bias.

2.1 Preliminary

Given an input pair (x_s, y_{mt}), the final output y_{final} is produced through our framework. We denote the English source sentence as x_s, the initial LLM translation in target language ℓ (Korean (Ko), Chinese (Zh), Japanese (Ja), German (De), or French (Fr)) as y_{mt}, and the final output of MELT as y_{final}, defined as follows:

$$\text{MELT} : (x_s, y_{\text{mt}}) \mapsto y_{\text{final}} \tag{1}$$

Translationese manifests in five patterns, namely faithfulness, voice, register, naturalness, and surface form. The meaning of each sentence must be preserved via bidirectional NLI entailment, and naturalness must be improved over the baseline y_{mt}. Any candidate that fails to satisfy either constraint is rolled back, and y_{mt} is returned as the output.

2.2 Detecting Translationese Patterns

Translationese can be divided into deterministic signals from surface form (Wang et al., 2023), such as comma overuse and em-dash frequency, and indeterministic signals from semantic and discourse categories, such as English-style introductory phrases and meta-commentary. The MELT detector performs two detection paths in parallel and integrates the results. A contiguous segment suspected of translationese within the initial translation y_{mt} is defined as a span $s = (b,\ e,\ c)$. b and e denote character offsets in y_{mt} ($0 \leq b < e \leq |y_{\text{mt}}|$), c denotes one of the categories in the translationese taxonomy. The detector produces a set of such spans, which serve as conditioning signals specifying *where and by which category criterion to edit* in the editing stage, and spans belonging to categories with a high risk of self-bias are subject to judgment by the debate gate.

The deterministic detector Φ_{surf} measures deviations of lexical markers relative to native corpus statistics, such as comma z-score and em-dash frequency, and extracts positions exceeding a threshold as a span set $\mathcal{S}_{\text{surf}}$ for the surface form and quantitative naturalness categories. Here, the z-score is the standardized value of marker frequency against the native corpus ($z = (r - \mu_{\mathcal{N}})/\sigma_{\mathcal{N}}$, where r is the marker frequency of the input and $\mathcal{N}$ denotes the native corpus), and a value of $z > \tau_z$ (default 2.0) is flagged as over-use, following common practices in stylometric authorship analysis (Patchala and Bhatnagar, 2018). The prompt-conditioned LLM detector Φ_{llm} receives the translationese taxonomy definition as a system prompt and outputs a span set $\mathcal{S}_{\text{llm}}$ for the faithfulness, voice, register, and naturalness categories. The two results are deduplicated using an Intersection over Union (IoU) threshold of τ_{IoU}=0.5, while spans with different categories

are retained to enable multi-category diagnosis:

$$\mathcal{S} = \mathcal{S}_{\text{surf}} \cup \{\, s \in \mathcal{S}_{\text{llm}} : \text{Keep}(s) \}. \quad (2)$$

Where $\text{Keep}(s)$ denotes the case in which the IoU with any span in $\mathcal{S}_{\text{surf}}$ is below τ_{IoU} or the category differs.

2.3 Span-Conditioned Editing

Based on the detected spans, MELT performs editing in two stages. First, a surface fixer T is applied to remove surface noise through meaning-preserving deterministic transformations, including Unicode NFC normalization, leading and trailing whitespace removal, and half-width and full-width unification for Zh, and the output is defined as $y' = T(y_{\text{mt}})$. Surface form corrections with potential for meaning change, such as comma and em-dash, are handled in the subsequent stage. The LLM editor then performs span-conditioned generation by injecting the category labels of the detected spans into the prompt:

$$y_e = \mathcal{E}\big(x_s,\, y',\, \mathcal{S}\big). \quad (3)$$

The prompt is composed of native instructions in the target language ℓ to mitigate prompt-level bias from English-based reasoning, and is provided together with the list of detected spans per category, such as "[naturalness] rewrite English-style passive constructions into active voice".

2.4 Verifying Edits

Directly applying the output of the LLM editor may introduce meaning distortion or hallucination. To prevent this, MELT reflects the edited result only when the candidate y_e satisfies all three complementary verification gates. First, the deterministic gate G_{surf} checks that the change in the metric set $\mathcal{M}$, including comma z-score, em-dash frequency, length variance, and epistemic-marker drift, does not deteriorate beyond the allowed range:

$$G_{\text{surf}}(y_e) = \mathbf{1}\big[\, \forall m \in \mathcal{M}, \; m(y_e) - m(y_{\text{mt}}) \geq -\tau_{\text{deg}} \,\big]. \quad (4)$$

where τ_{deg}=0.5 is a threshold set to alleviate excessive rejection observed on a pilot subset. To verify meaning preservation, bidirectional NLI entailment (Maynez et al., 2020) is applied. Forward p_f and backward p_b entailment probabilities are estimated using mDeBERTa-XNLI, and the gate is passed only when the entailment scores in both directions meet or exceed the threshold:

$$G_{\text{nli}}(y_e) = \mathbf{1}\big[\, \min(p_f, p_b) \geq \tau_{\text{nli}}^{(\ell)} \,\big]. \quad (5)$$

Bidirectional verification simultaneously detects omission (where only p_f is high) and hallucination (where only p_b is high), and the threshold is calibrated per language. Finally, the LLM-based universal judge $\mathcal{J}_u$ evaluates factual preservation of the edited output y_e:

$$G_{\text{judge}}(y_e) = \mathcal{J}_u\big(x_s, y_{\text{mt}}, y_e\big) \in \{0, 1\}. \quad (6)$$

This stage additionally evaluates factual consistency, register consistency, and discourse naturalness that are difficult to capture through deterministic metrics and NLI-based verification alone.

The edited result is reflected only when all three verification gates return True; otherwise, the original translation y_{mt} is retained. Here, $\mathcal{G} = \{G_{\text{surf}}, G_{\text{nli}}, G_{\text{judge}}\}$ is as follows:

$$y_{\text{final}}^{(\text{verify})} = \begin{cases} y_e & \text{if } \prod_{g \in \mathcal{G}} g(y_e) = 1, \\ y_{\text{mt}} & \text{otherwise.} \end{cases} \quad (7)$$

2.5 Debate on Self-bias-prone Categories

Relying solely on a universal judge is vulnerable to the self-bias problem, in which an LLM tends to regard its own pretraining distribution as *natural expression* (Xu et al., 2024; Chen et al., 2025). This problem does not manifest uniformly across all categories. While surface form errors and faithfulness categories can be complemented by surface metrics and NLI-based verification, categories related to lexical naturalness and voice rely heavily on native-level linguistic intuition, making stable judgment difficult with a universal judge alone. Accordingly, MELT applies an additional debate gate exclusively to categories with a high risk of self-bias. *Voice anchor* refers to stylistic characteristics that should be preserved to maintain the speaker's characteristic voice, such as distinctive lexical choices or sentence patterns, whereas *voice drift* refers to changes that move the output away from these characteristics toward more generic or model-like expression.

$$\mathcal{D}_{\text{sel}} = \{\, \text{lexical naturalness}, \; \text{voice anchor}, \; \text{voice drift} \,\}. \quad (8)$$

This limits additional inference costs by performing debate only for the categories that require it.

The debate consists of two stages. First, the universal judge $\mathcal{J}_u$ and the native judge $\mathcal{J}_n^{(\ell)}$ each

produce a verdict and a confidence score for the same input. If the verdicts of the two judges agree and the difference in confidence is small, the judgment is finalized. If the two verdicts conflict, a cross-examination stage is performed, in which each judge re-evaluates by referring to the verdict and reasoning of the other judge. Through this process, the universal judge and the native judge mutually compensate for each other's blind spots.

For each category $d \in \mathcal{D}_{\text{sel}}$, the confidences of the two judges are combined using a native priority weight $w_n^{(d)} \in [0, 1]$. The debate gate is passed when the weighted sum of confidences favoring the edit $\hat{c}_d$ meets or exceeds the threshold:

$$G_{\text{debate}}^{(d)}(y_e) = \mathbf{1}\left[\, \hat{c}_d \geq \tau_{\text{debate}} \,\right]. \tag{9}$$

Since lexical naturalness depends on native lexical preference, and voice anchor and voice drift are centered on author-specific register consistency, the weight of the native judge is set relatively higher for these categories.

2.6 Final Verification and Decision

The final output of MELT is determined by applying the verification gates (Section 2.4) and the selective debate gates (Section 2.5):

$$y_{\text{final}} = \begin{cases} y_e & \text{if all gates pass,} \\ y_{\text{mt}} & \text{otherwise.} \end{cases} \tag{10}$$

The edited result y_e is reflected only when all verification gates ($G_{\text{surf}}, G_{\text{nli}}, G_{\text{judge}}$) and selective debate gates ($G_{\text{debate}}^{(d)}$, $d \in \mathcal{D}_{\text{sel}}$) return 1; otherwise, the original translation y_{mt} is retained. Through this process, MELT operates in four stages of detect, edit, verify, and debate without additional model training or weight updates, preventing edited results that fail verification from being reflected in the final output.

3 Experiments

3.1 Setup

Datasets For the multilingual experiments, we use translation tasks based on FLORES-200 (NLLB Team et al., 2022), covering five language pairs with English as the source language and Korean (Ko), Chinese (Zh), Japanese (Ja), German (De), and French (Fr) as target languages. For the monolingual experiments, we evaluate texts generated by back-translating texts written by authors in each language through English using an LLM. See Appendix B for details.

Models We use GPT-4o-mini as the editor for all methods. Claude-haiku 4.5 is used as the universal judge, while Gemini-2.5-Pro and GPT-5 are included for cross-family robustness checks. Averaging PE across judges from different model families mitigates model-specific preference effects and yields a more stable evaluation signal, consistent with recent findings on self-preference bias and multi-judge evaluation in LLM-as-a-Judge systems (Wataoka et al., 2025; Verga et al., 2024). We further use the PE gap between the same-family judge (`GPT`) and cross-family judges (`Claude` / `Gemini`) as a direct proxy for quantifying self-bias. For the debate gate, we use EXAONE-4.0-32B as the native judge for Korean and Qwen3-32B for Chinese, Japanese, German, and French, while reference-free metrics are computed with Qwen2.5-7B for PPL and XLM-RoBERTa-large for BERTScore and MAUVE. See Appendix F.

Baseline models. We compare against three representative editing baselines. *Naive Polish* performs a single polishing pass using a generic instruction without any explicit detection or verification stage. *LiL Polish* reproduces the polishing prompt proposed in Lost in Literalism (Li et al., 2025), consisting of a single polishing call that emphasizes native-like phrasing and stylistic adaptation. *Self-Refine* (Madaan et al., 2023) adopts an iterative refinement framework that alternates between feedback generation and refinement, terminating when no further improvement is suggested or after a maximum of three refinement iterations.

3.2 Evaluation Metrics

We evaluate performance from three perspectives: preference, distribution matching, and lossless. See Appendix C for details.

Preference We use pairwise LLM judgments to compute the prefer-edit rate (PE) as the primary evaluation metric. PE is defined as the proportion of cases in which the judge considers the method output y_p more natural than the baseline output y_o:

$$\text{PE} = \frac{1}{n} \sum_i \mathbf{1}\left[L_p^{(i)} - L_o^{(i)} \geq 0.5\right], \tag{11}$$

where $L^{(i)}$ denotes the average Likert score over the nine rubrics. To mitigate position bias, we conduct two evaluations by swapping the presentation order of (y_o, y_p), and report their average. See Appendix C.2 for details.

Distribution matching To measure distributional similarity between method outputs and human references, we use MAUVE (Pillutla et al., 2021) and calibration error, defined as $\Delta_{\mathrm{O}} = \mathrm{PE} - \mathrm{Oracle\ PE}$, where Oracle PE denotes the PE of human references. As auxiliary metrics, we report automatic indicators of translation naturalness, including lexical density (Lex), length variance (Len), and perplexity (PPL) (Li et al., 2025), along with stylometric distance metrics such as BERTScore, cosine similarity, lexical JSD, and comma and em-dash frequency differences.

Lossless To evaluate semantic loss during editing, we measure the bidirectional NLI entailment rate (Maynez et al., 2020). The threshold $\tau_{\mathrm{nli}}^{(\ell)}$ is calibrated for each language using the 95th percentile of the entailment score distribution.

3.3 Multilingual Main Results

Table 2 presents PE and MAUVE for three prior baselines (Naive Polish, LiL Polish, Self-Refine) and three MELT variants: *MELT-LLM*, which uses a single LLM for detect–edit–verify without rule-based verification or multi-agent debate; *MELT-Hybrid*, which additionally incorporates a rule-based surface detection and verification layer; and *MELT-MAS*, which further adds selective multi-agent debate for self-bias-prone categories. The initial LLM translation (raw MT) serves as the baseline across five target languages on FLORES-200. PE is averaged over three judges (Claude-Haiku 4.5, Gemini-2.5-Pro, and GPT-5), and $\overline{|\Delta_{\mathrm{O}}|}$ quantifies calibration error as the average absolute difference between PE and Oracle PE across the five languages, where lower values indicate closer calibration to the human reference distribution.

The experimental results show that MELT simultaneously achieves calibration and meaning preservation in multilingual translationese mitigation. 1) MELT variants consistently maintain results closest to Oracle preference, exhibiting more stable post-editing characteristics compared to Naive Polish or Self-Refine. In contrast, some baselines show a tendency to substantially exceed Oracle preference despite recording high PE, which we interpret as the effect of self-preference bias, where LLM judges favor LLM-style expressions closer to their pretraining distribution over human-like naturalness. 2) Self-Refine shows lower preference than the baseline across several languages, suggesting that iterative rewriting can induce over-editing that causes outputs to deviate from the original meaning and stylistic distribution. MELT controls edits with potential for meaning distortion through its detect–edit–verify structure, verification gates, and rollback mechanism, enabling stable editing even in a training-free setting. In particular, MELT variants show notably larger improvements in languages with wordy structures and register variation, such as Japanese, indicating that translationese is a language-universal phenomenon while exhibiting different surface realizations across languages. These results demonstrate that the unified training-free framework combining typology-aware category detection and selective verification can operate as a consistent post-editing strategy across language types.

Moreover, MELT-LLM achieves the lowest calibration error among all baselines ($\overline{|\Delta_O|} = 3.9$), while requiring only 3.0 model calls per sample. As shown in Table 10, MELT-LLM and MELT-Hybrid add only one model call compared to the single-pass baselines, resulting in an approximately 2-second increase in latency. In contrast, MELT-MAS incurs substantially higher latency due to native-judge inference. Compared with Self-Refine (6.9 model calls per sample, $\overline{|\Delta_O|} = 15.5$), these results suggest that the performance gains are not simply due to increasing the number of model calls, but primarily stem from the proposed detect–edit–verify formulation itself.

3.4 Monolingual Main Results

Table 3 presents the main results in the monolingual author-style recovery scenario. The experimental setup involves translating pre-LLM-era authors' original texts into English and then back-translating them into the source language, a process that tends to weaken the original author's idiolect and register and converge toward a more averaged LLM-style expression. Each method takes the back-translated output as input and attempts to recover the original author's style, and performance is measured using pairwise preference-based PE and BERTScore F1 against the author's original text.

The experimental results reveal different patterns in the monolingual recovery setting compared to the multilingual translation setting. LiL Polish records the highest PE in Korean, Chinese, German, and French, while MELT variants show superior performance in Japanese. Meanwhile, the PE gap between Oracle and each method is relatively small in Japanese, German, and French, which we in-

Method	Ko		Zh		Ja		De		Fr		
	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	$\overline{\lvert\Delta_O\rvert}$↓
Baselines (prior work)											
Naive Polish	23.8	.887	32.9	.928	29.3	.947	43.6	.966	45.7	**.986**	6.8
LiL Polish (Li et al., 2025)	27.0	.902	35.6	.901	34.1	.928	**47.4**	.971	**48.1**	.984	5.6
Self-Refine (Madaan et al., 2023)	20.8	.903	17.1	.906	29.2	.946	30.4	.967	23.0	.979	15.5
Ours											
MELT-LLM	**29.8**	**.937**	**36.5**	.895	41.8	.902	43.8	.950	44.4	.973	**3.9**
MELT-Hybrid	27.6	.908	21.5	**.930**	**42.6**	**.932**	41.8	.962	40.7	.978	6.4
MELT-MAS	25.9	.933	20.4	.908	42.3	.929	41.6	**.973**	41.1	.979	6.6
Oracle (human reference)	27.7	—	41.5	—	43.0	—	42.8	—	40.1	—	—

Table 2: Main results across five FLORES-200 target languages. The best PE per language is shown in **bold**, and the value closest to Oracle is underlined. LiL Polish is a training-free reproduction that applies only the prompt from Li et al. (2025) to GPT-4o-mini.

Method	Ko		Zh		Ja		De		Fr	
	PE↑	BS↑	PE↑	BS↑	PE↑	BS↑	PE↑	BS↑	PE↑	BS↑
Baselines (prior work)										
Naive Polish	4.8	**.892**	14.6	**.903**	11.0	.878	29.8	.960	32.3	.930
LiL Polish (Li et al., 2025)	**11.6**	.891	**36.3**	.903	25.3	.878	**37.2**	.960	**36.7**	.928
Self-Refine (Madaan et al., 2023)	2.4	**.892**	9.4	.902	12.8	.877	11.8	.958	14.8	.934
Ours										
MELT-LLM	8.3	.891	8.4	.902	24.0	.877	19.0	.957	27.1	.931
MELT-Hybrid	7.0	.891	5.9	.902	28.5	.877	18.8	.958	27.8	.931
MELT-MAS	7.3	.891	5.9	.902	**28.7**	.877	15.8	.959	32.3	.932
Oracle (author original)	29.6	—	70.4	—	75.5	—	24.0	—	23.5	—

Table 3: Monolingual author-style recovery main results. Pipeline: pre-LLM-era author original → English forward MT → target language back MT → recovery attempt by each method. PE denotes prefer-edit rate and BS denotes BERTScore F1. Full metric set and per-judge breakdown are provided in Appendix Table 15.

terpret as the LLM judge favoring more generic and standardized MT-style expressions over outputs that preserve pre-LLM-era writing styles. This suggests that LLM judge evaluation alone is insufficient to fully reflect human stylistic authenticity, confirming the necessity of calibration-aware evaluation. BERTScore remains at nearly identical levels across methods and languages, confirming that discriminability appears along the stylistic preference axis rather than semantic similarity. Full metric sets and per-judge breakdowns are provided in Appendix Table 15.

4 Results and Analysis

We analyze whether the key design components of MELT perform their intended roles in practice. The analysis examines the effects of the verification gate and selective debate on calibration and meaning preservation, and compares self-bias patterns across languages under different cross-family judge configurations. Finally, through monolingual author-style recovery and reference-free naturalness analysis, we evaluate whether MELT contributes to stylistic authenticity and idiolect restoration beyond fluency polishing.

4.1 Cross-family Judge Composition and Self-bias

Table 4 and Figure 3 show that self-preference bias in the LLM-as-a-judge setting varies substantially depending on the judge family. Evaluating the same GPT-4o-mini editor output while varying only the judge, Oracle PE is lowest for the same-family judge GPT-5 and tends to increase sequentially for cross-family judges Claude-haiku 4.5 and Gemini-2.5-Pro. GPT-5 evaluates GPT-4o-mini MT outputs as more natural than human references in Korean, Chinese, and Japanese, which can be interpreted as self-preference bias where same-family judges favor expressions similar to their own generation distribution. Cross-family judges maintain relatively stable preference for human reference, indicating that cross-family calibration is necessary for multilingual post-editing evaluation. Across languages, Korean shows the lowest Oracle preference across all judges, which we attribute to the conciseness

Judge	Ko	Zh	Ja	De	Fr	Avg	$\overline{\Delta}_{\text{lik}}$
GPT-5	20.8	28.5	31.2	36.4	36.6	30.7	−0.09
Haiku 4.5	25.3	45.8	44.5	43.0	40.9	39.9	+0.01
Gemini 2.5 Pro	**36.8**	**50.3**	**53.2**	**49.0**	**42.8**	**46.4**	**+0.09**

Table 4: Judge self-bias, per family. Per-language performance is a decomposition of Oracle PE from Table 2 by LLM. Lower values indicate stronger self-bias, where the judge prefers same-family MT output over human references. Per-rubric Likert metrics and scores are in Appendix C.1 and Appendix J.

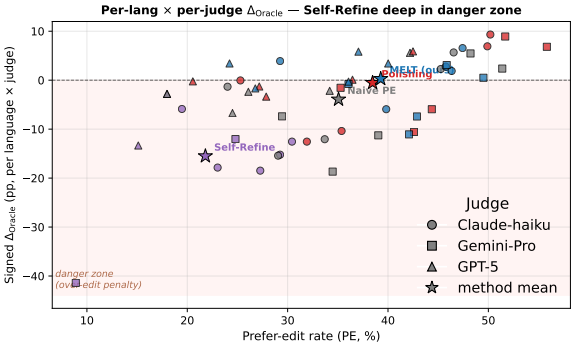

Figure 3: Per-language × per-judge signed Δ_{O} scatter. Each method is represented by 15 points (5 languages × judges), where ⋆ denotes the method mean. The negative Δ_{O} region indicates a danger zone where the judge rejects edits more than Oracle.

and implicature of professionally translated Korean not fully aligning with the fluency prior of current LLMs. In Figure 3, MELT variants are distributed relatively stably near Oracle, demonstrating that the calibration-aware detect-edit-verify structure is more robust to changes in judge family.

Although LiL Polish achieves the highest raw PE in German and French, these scores exhibit larger deviations from the corresponding Oracle preference, resulting in increased calibration error. Moreover, the PE differences between LiL Polish and MELT are not statistically significant in either language (Appendix H). In contrast, MELT consistently remains closer to the Oracle preference across all languages, indicating better calibration to human-reference preferences.

4.2 Reference-free Naturalness

Separate from the LLM judge-based preference and stylometric divergence analysis, we analyze the output of each method using the reference-free naturalness metrics, such as lexical (Lex), length (Len), and PPL, proposed by Li et al. (2025) (Appendix C.4). These metrics measure how well output texts align with the natural language distribution, but can simultaneously be interpreted as LM-aligned fluency metrics that reflect similarity to the training distribution of modern LLMs.

Table 5 shows that the PPL of pre-LLM-era authors' original texts (Original) is consistently higher than that of all post-editing outputs. Human-authored original texts are evaluated as unnatural by PPL, as PPL measures alignment with the modern LM distribution rather than human-likeness itself. In practice, most post-editing methods converge toward fluent and averaged LM-style expressions, with LiL Polish recording the lowest PPL and most stable length variance across several languages. This tendency is observed equally in the FLORES translation experiments, indicating that polishing-based instructions produce outputs most closely aligned with the modern LLM prior.

In contrast, MELT variants show a tendency to shift toward the author original distribution in the Lex (content-word ratio) metric across multiple languages. This indicates that MELT does not perform fluent rewriting, but operates in a way that better preserves content-word distribution and stylistic markers. This difference is pronounced in languages where discourse convention and register variation have a strong influence, such as Korean and Japanese. High performance on LiL metrics does not imply human-like stylistic authenticity or author-style preservation, and we confirm that MELT is closer to stylistic distribution recovery than LM-style fluency optimization.

5 Conclusion

In this paper, we propose MELT, a training-free method that mitigates translationese across five languages and brings generated text closer to native naturalness. Our empirical analysis of MELT focuses on three aspects: 1) For *Typology-aware detection*, we classify translationese patterns into five categories for each target language and quantitatively measure their typology-conditioned realizations. 2) For *Training-free pipeline*, we show that MELT can mitigate translationese without weight updates while preserving meaning. The experimental results on FLORES-200 show that MELT-LLM achieves the lowest average calibration error across the three judges among all baselines, with $\overline{|\Delta_{\text{O}}|} = 3.9$ pp. MELT-Hybrid shows the closest results to Oracle PE in four of the five languages. Although LiL Polish attains the highest raw preference scores in German and French, MELT con-

Method	Ko			Zh			Ja			De			Fr		
	Lex↑	Len	PPL↓	Lex↑	Len	PPL↓	Lex↑	Len	PPL↓	Lex↑	Len	PPL↓	Lex↑	Len	PPL↓
MT round-trip	0.604	1.697	10.2	0.503	1.420	30.3	0.416	1.417	15.9	0.490	1.029	12.7	0.412	1.063	13.3
Naive Polish	0.605	1.695	10.2	0.505	1.411	29.7	0.417	1.413	15.8	0.488	1.033	12.4	0.415	1.060	12.6
LiL Polish (Li et al., 2025)	0.605	1.793	10.0	0.506	1.416	28.8	0.418	1.479	15.3	0.487	1.043	12.3	0.413	1.071	12.2
Self-Refine (Madaan et al., 2023)	0.606	1.675	10.2	0.504	1.408	30.3	0.416	1.413	15.8	0.489	1.021	12.6	0.414	1.059	12.9
MELT-LLM	**0.608**	1.681	10.3	**0.507**	1.414	30.8	0.429	1.365	16.7	0.493	1.039	12.4	**0.418**	1.054	12.9
MELT-Hybrid	**0.608**	1.666	10.2	0.506	1.419	30.7	0.430	1.378	16.9	0.492	1.030	12.4	**0.418**	1.054	12.6
MELT-MAS	0.607	1.688	10.2	0.506	1.431	30.3	**0.431**	1.375	16.9	**0.494**	1.031	12.6	0.417	1.051	12.7
Original (pre-LLM author)	*0.653*	*1.000*	*24.6*	*0.501*	*1.000*	*23.0*	*0.447*	*1.000*	*31.1*	*0.501*	*1.000*	*14.0*	*0.233*	*1.000*	*20.8*

Table 5: LiL-style naturalness on the author-recovery experiment (monolingual setup). 5 languages × 7 methods × approximately 200 paragraphs; each method output is compared against the *author's original text*. Metric details are in Appendix C.4.

sistently remains better calibrated to the Oracle preference across languages. 3) for *Idiolect restoration*, we confirm that, in a monolingual recovery scenario involving 49 pre-LLM-era authors, MELT better preserves BERTScore and stylometric distributions in the direction of the original authors than the baselines.

In future work, we plan to extend the five category taxonomy proposed in this paper to a broader range of linguistic patterns. We will also expand native-speaker validation to cover more pattern types and examine settings where translationese accumulates across multiple stages to further assess the generalizability of MELT. Building on these extensions, we will explore the integration of MELT as a post-editing component within broader agentic translation workflows.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education RS-2026-25471023, RS-2025-00523385. This research was supported by the ANCHOR program through the Gangwon ANCHOR Center, funded by the Ministry of Education (MOE) and the Gangwon State (G.S.), Republic of Korea. (2026-ANCHOR-10-006)

Limitations

The five category taxonomy and detect-edit-verify pipeline of MELT are designed to target *LLM-specific unnaturalness*, including translationese, post-training distortion, and English-structure mimicry. General copy-editing errors, such as spelling mistakes, subject-predicate agreement, or clause-comma placement, are not primary target. Surface errors may also be addressed during category-conditioned editing by the LLM editor in the edit gate. Cases that involve semantic distortion are blocked by the bidirectional NLI gate in the verify gate. However, we do not report quantitative metrics such as "spelling accuracy" or "grammar error rate". MELT is not a replacement for a grammar checker. For production use, we recommend applying MELT followed by a separate spell or grammar checker.

MELT modifies stylistic characteristics while constraining changes to semantic content through the NLI gate, but this capability could potentially be misused to evade MT detection or authorship analysis. We do not overstate the effectiveness of existing defenses. Stylometric features alone may not reliably distinguish human-written text from machine-generated text, and MELT explicitly reduces the very signals targeted by such analyses. Therefore, rather than claiming that our analysis provides a countermeasure, we identify the reliable detection of MT outputs edited by MELT as an open challenge for future work. In addition, our taxonomy and native-speaker judges implicitly assume a language-specific notion of *nativeness*, although perceived naturalness can vary across dialects, generations, and registers.

Ethical Considerations

All datasets and models used in this work are publicly available, and pre-LLM-era literary texts used in the monolingual author-style recovery experiments are sourced from public domain repositories, including Aozora Bunko, Wikisource, and Project Gutenberg. We are not aware of any direct negative societal impacts specific to this work beyond those generally associated with the advancement of large language models.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*, pages 280–290, Online. Association for Computational Linguistics.

Tom B. Brown and 1 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2018. Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. EditEval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331*.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning (ICML)*.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. *Preprint*, arXiv:2309.16349.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your large language models are leaving fingerprints. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 85–95. Association for Computational Linguistics.

Shervin Minaee and 1 others. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *Preprint*, arXiv:2309.05196.

Jagadeesh Patchala and Raj Bhatnagar. 2018. Authorship attribution by consensus among multiple features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2766–2777, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Libo Qin and 1 others. 2025. Large language models meet nlp: A survey. *Preprint*, arXiv:2405.12819.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7737–7746, Online. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024a. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling (COLM)*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024b. A long way to go: Investigating length correlations in rlhf. *Preprint*, arXiv:2310.03716.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play favorites: A statistical method to measure self-bias in llm-as-a-judge. *Preprint*, arXiv:2508.06709.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-preference bias in llm-as-a-judge. *Preprint*, arXiv:2410.21819.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Wayne Xin Zhao and 1 others. 2026. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Yutao Zhu and 1 others. 2025. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54.

Appendix

Contents

A Related Work 13

B Datasets 13

C Evaluation Metrics in Detail 14

D Translationese Taxonomy in Details 16

E Hyperparameters 17

F Model Cards 17

G Per-language Metric Breakdown 17

H Statistical Significance Analysis 17

I Calibration Metrics 18

J Per-rubric Mean Likert Score 18

K Judge-Free Stylometric Analysis of Translationese Recovery 18

L Qualitative MELT post-editing examples 21

M Cross-family Editor Evaluation 23

A Related Work

Measuring and Mitigating Translationese Translationese analysis has examined the unnaturalness of Neural Machine Translation(NMT) from multiple perspectives since Volansky et al. (2015) characterized translationese through automatic classification based on statistical features. Prior studies have measured various aspects of translationese, including reduced lexical diversity (Vanmassenhove et al., 2019), the acceleration of post-editese (Toral, 2019), and the reassessment of human-parity claims (Toral et al., 2018). These studies have shown that translationese is not a mere stylistic deviation, but a systematic phenomenon observable across statistical features and syntactic choices (Daems et al., 2018). Approaches for distinguishing human and machine translation based on information theory (Bizzoni et al., 2020) and translationese aware multilingual NMT (Riley et al., 2020) have also been proposed as methods for measuring and mitigating translationese. Translationese has been shown to affect not only machine translation itself but also evaluation sets, with test sets containing translationese reported to distort system comparisons (Zhang and Toral, 2019). Jalota et al. (2023) propose a translation-based style transfer approach that mitigates translationese without relying on parallel data. As LLMs have become widely adopted, Li et al. (2025) mitigate unnaturalness through SFT data curation based on reference polishing and unnatural-instance filtering, while Guo et al. (2025) address multilingual unnaturalness using corpus-level lexical JSD measurements and DPO alignment. However, these mitigation approaches rely on training-time methods such as SFT or DPO, thereby incurring additional retraining costs.

Characterizing Post-training Distortion Post-training distortion analysis addresses non-translationese aspects of translationese. Singhal et al. (2024a) show that reward improvements from RLHF can be reduced to response length inflation, while Hosking et al. (2024) analyze how a single preference score can conflate factuality with assertiveness. Sharma et al. (2024) demonstrate that sycophancy originates from preference data. From the perspective of output diversity, Kirk et al. (2024) and Padmakumar and He (2024) show that RLHF and writing with language models can reduce output diversity. Liang et al. (2024) and McGovern et al. (2025) quantify lexical fingerprints characteristic of LLM outputs, including words such as "delve" and "intricate".

However, these studies mainly focus on diagnosis and characterization, and there has been no multilingual attempt to mitigate both translationese and post-training distortion within a unified inference-time framework. We integrate translationese and post-training distortion under a five category taxonomy and propose a training-free post-editing approach for mitigating translationese.

LLM-as-a-Judge and Limitations LLM-as-a-judge has been widely adopted as an efficient means of replacing or assisting human judgment in large-scale generation evaluation. Zheng et al. (2023a) demonstrate that LLM judges can serve as useful comparative evaluation tools on MT-Bench and Chatbot Arena, and the potential of LLM judges for translation quality evaluation has also been reported (Kocmi and Federmann, 2023). However, several studies show that LLM judges are not always neutral or stable. Xu et al. (2024) show a tendency toward self-bias, while Wataoka et al. (2025) and Chen et al. (2025) quantitatively analyze this bias as a potential structural characteristic of judges, and Spiliopoulou et al. (2025) propose a statistical method for measuring self-bias. Hosking et al. (2024) analyze the limitations of single preference scores, and Singhal et al. (2024b) show that length inflation following RLHF can lead to overestimation of fluent-but-long outputs. To mitigate these issues, we propose an approach that combines not only an LLM judge but also surface-based verification and bidirectional NLI.

B Datasets

Table 6 summarizes the sample size, average length, and sources of the two corpora used in this paper. We then describe the selection rationale and composition of each corpus in detail.

Multilingual setup: FLORES-200. We use 1,012 paired sentences from the devtest split of NLLB Team et al. (2022), with English as the source language and five target languages: Ko `kor_Hang`, Ch Simplified `zho_Hans`, Ja `jpn_Jpan`, De `deu_Latn`, Fr `fra_Latn`. For Ko, only 1,010 samples are valid (2 excluded due to NLI tokenizer overflow), and Oracle (human reference) evaluation for Fr is based on 1,008 samples (4 references are empty). All samples are at the sentence level, with average length ranging from 70 (Ko) to 150

(a) FLORES-200 devtest: main evaluation

Lang	N_{src}	N_{ref}	chars	words
En (source)	1012	—	130.4	21.6
Ko	1012	1010	65.1	15.2
Zh	1012	1012	42.7	—
Ja	1012	1012	56.3	—
De	1012	1012	152.0	21.5
Fr	1012	1008	155.7	25.3

(b) Pre-LLM-era author corpus: Frame B (idiolect)

Lang	Auth.	Para.	chars	Source
Ko	10	200	351.9	Wikisource
Zh	8	135†	354.3	Wikisource
Ja	10	200	362.1	Aozora Bunko
De	10	200	320.3	Wikisource
Fr	11	220	330.6	Gutenberg
Total		**955**		

Table 6: Dataset statistics. (a) FLORES-200 (NLLB Team et al., 2022) devtest split. N_{src} = paired source sentences, N_{ref} = valid human reference targets (Ko: 2 NLI-tokenizer overflows; Fr: 4 empty references). *chars* = mean character length per sentence; *words* = mean whitespace-split token count (omitted for Zh/Ja which lack inter-word spaces). English source is identical across all 5 target conditions. (b) Pre-LLM-era author corpus used for Frame B (idiolect preservation), restricted to authors active before any LLM training cutoff (pre-2018 publication) so the baseline is "LLM-uncontaminated." Per paragraph filtered to 80–400 characters; multiple works per author are sampled to inject within-author topic diversity. †Zh: 17 paragraphs identified as chapter title lists / table-of-contents (rather than prose) were excluded from author-recovery evaluation (152 collected → 135 used).

(De) characters per language.

Monolingual setup: Pre-LLM-era authors corpus. We collected 972 paragraphs from 49 authors across five languages, targeting approximately 20 paragraphs per author. For Zh, some samples corresponded to chapter-list TOC; after removing 17 such samples, 955 paragraphs were used for the author-recovery evaluation.

All authors were active prior to the LLM training cutoff (2018–2023), serving as an "LLM-uncontaminated" baseline. Each paragraph is filtered to the range of 80–400 characters, and samples are drawn from multiple works per author to ensure within-author topic diversity. Sources and author list:

- Ja (10 authors, 200 paragraphs), Aozora Bunko[1]: Soseki Natsume (夏目漱石), Mori Ogai (森外), Ryunosuke Akutagawa (芥川之介), Osamu Dazai (太宰治), Kenji Miyazawa (宮賢治), Izumi Kyoka (泉鏡花), Higuchi Ichiyo (口一葉), Sakaguchi Ango (坂口安吾), Yokomitsu Riichi (光利一), Nagai Kafu (永井荷風).
- Ko (10 authors, 200 paragraphs), Korean Wikisource[2]: Dongin Kim (김동인), Jingeon Hyeon (현진건), Sang Lee (이상), Hyoseok Lee (이효석), Mansik Chae (채만식), Gwangsu Lee (이광수), Hun Sim (심훈), Yujeong Kim (김유정), Seohae Choi (최서해), Dohyang Na (나도향).
- Zh (8 authors, 152→135 paragraphs after TOC filtering), Chinese Wikisource[3] (Republican-era 작가): Yu Dafu (郁達夫), Zhu Ziqing (朱自), Qu Qiubai (瞿秋白), Hu Shi (胡適), Liang Qichao (梁超), Lao She (老舍), Zhou Zuoren (周作人), Lu Xun (魯迅).
- De (10 authors, 200 paragraphs), German Wikisource[4]: Johann Wolfgang von Goethe, Friedrich Schiller, Thomas Mann, Rainer Maria Rilke, Hugo von Hofmannsthal, Arthur Schnitzler, Theodor Storm, Theodor Fontane, Heinrich von Kleist, Adalbert Stifter.
- Fr (11 authors, 220 paragraphs), Project Gutenberg[5]: George Sand, Victor Hugo, Gustave Flaubert, Guy de Maupassant, Émile Zola, Honoré de Balzac, Marcel Proust, Stendhal, Charles Baudelaire, Alphonse Daudet, Paul Verlaine.

C Evaluation Metrics in Detail

C.1 Per-rubric Likert mean

We evaluate post-edited outputs using a 9-rubric LLM judge. For each sample, the judge scores each candidate output on a 1–5 Likert scale along nine dimensions: faithfulness, adequacy, fluency, syntax, lexicon, discourse, register, conciseness, and human-likeness. The resulting per-rubric scores are used both as auxiliary diagnostic metrics and as the basis for pairwise preference evaluation.

Claude-haiku-4.5 is used as the default judge, and the default judge is selected from a different model family than the editor to mitigate self-bias.

[1] https://huggingface.co/datasets/globis-university/aozorabunko-clean
[2] https://ko.wikisource.org/
[3] https://zh.wikisource.org/
[4] https://de.wikisource.org/
[5] https://www.gutenberg.org/

To analyze judge-family dependence, evaluation results using Gemini-2.5-Pro and GPT-5 are also reported in the appendix. We report corpus-level means for each of the nine rubrics. Results are shown in Tables 21, 22, and 23.

C.2 Prefer-edit rate (PE)

LLM judge-based pairwise preference has been studied in the context of human pairwise preference in RLHF (Stiennon et al., 2020), and has been extended to the LLM-as-a-judge framework (Zheng et al., 2023b) and automated win-rate evaluators (Dubois et al., 2023) as a representative evaluation method for open-ended text generation. We adapt this approach to the multilingual post-editing setting, following a setup similar to edit-task pairwise evaluation (Dwivedi-Yu et al., 2022). Specifically, the LLM judge using nine rubrics compares the method output $y_p^{(i)}$ and the baseline output $y_o^{(i)}$ for sample i, and produces 1–5 Likert scores $L_p^{(i)}$ and $L_o^{(i)}$ for each candidate. To mitigate ordering bias, we perform position-swapped evaluation by exchanging the presentation order of (y_o, y_p) (Zheng et al., 2023b).

Recent studies show that LLM judges exhibit self-preference bias (Xu et al., 2024; Chen et al., 2025), where they favor LLM-style outputs similar to their pretraining distribution, as well as length bias toward longer responses (Dubois et al., 2024). Raw PE can simultaneously reflect both actual quality improvements and judge bias effects, and high PE does not necessarily imply improved naturalness as perceived by humans.

To address this, we follow the calibration perspective of Chen et al. (2025) and use human references as calibration anchors. In this setup, human references are not directly provided in the judge prompt; pairwise evaluation is performed using only the source sentence and two candidate outputs. Human references are used exclusively in the separate Oracle PE computation, serving as calibration anchors. Oracle PE denotes the prefer-edit rate obtained by the human reference under the same judge and baseline conditions. Method outputs and human references are thus evaluated under the same judge bias, and the difference between the two values represents a relative preference that partially cancels out the common bias of the judge.

$$\Delta_O^{(m,\ell)} = \mathrm{PE}^{(m,\ell)} - \mathrm{PE}^{(\mathrm{Oracle},\ell)},$$
$$\overline{|\Delta_O|}^{(m)} = \frac{1}{|\mathcal{L}|}\sum_{\ell}\left|\Delta_O^{(m,\ell)}\right|. \tag{12}$$

A smaller $|\Delta_O|$ value indicates that the method output is closer to the Oracle (human reference). When $\Delta_O > 0$ (over-shoot), the judge prefers the method output over the human reference, and this over-shoot phenomenon suggests that LLM-style characteristics targeted for removal may have instead been reinforced (Xu et al., 2024). We therefore use $\overline{|\Delta_O|}$ as the primary calibration metric in place of raw PE. $\overline{|\Delta_O|}$ in the last column of Table 2 is averaged over five languages. Per-language Δ_O values are provided in Appendix Table 13.

C.3 Auxiliary Metrics

Length-controlled PE (LC-PE). We apply the length-controlled evaluation of Dubois et al. (2024) to PE. For each language ℓ, the linear relationship between PE scores and output length is estimated as follows.

$$\mathrm{PE} = a + \beta \log(|y_e|/|y_{\mathrm{MT}}|) + \epsilon \tag{13}$$

$|y_e|/|y_{\mathrm{MT}}|$ denotes the length ratio of the post-edited output relative to the initial MT output, which is subtracted from the PE score to define length-controlled PE.

$$\text{LC-PE}^{(m,\ell)} = \mathrm{PE}^{(m,\ell)} - \hat{\beta}^{(\ell)} \log\left(\frac{|y_e^{(m,\ell)}|}{|y_{\mathrm{MT}}^{(\ell)}|}\right). \tag{14}$$

LC-PE is a PE score that mitigates length bias favoring verbose outputs. Results are provided in Appendix 13.

MAUVE. Following the distribution-matching metric of Pillutla et al. (2021), we measure the distributional difference between human reference P and method output Q. Both text sets are transformed into a contextual embedding space, discrete distributions are estimated through k-means quantization, and the KL-divergence frontier is computed with respect to the mixture distribution $M_\lambda = \lambda P + (1-\lambda)Q$.

$$C(P,Q) = (\mathrm{KL}(Q\|M_\lambda), \mathrm{KL}(P\|M_\lambda)) \tag{15}$$

The final score is a normalized value of the area under the frontier, ranging in $[0,1]$, where values closer to 1 indicate better alignment of the output

distribution with the human reference distribution. Since MAUVE evaluates the distributional alignment of the entire output set rather than similarity at the individual sentence level, both under-generation with insufficient diversity relative to the reference distribution and over-generation that deviates from the reference distribution are reflected as lower scores. Used in Tables 2 and 15.

C.4 LiL-style metrics (Lex / Len / PPL)

Following the automatic evaluation protocol of Li et al. (2025), we report three LiL-style naturalness metrics: lexical density (Lex), length variety (Len), and perplexity (PPL). These metrics are designed to capture surface-level signals of translationese and target-language naturalness.

Lex. Lex measures lexical density, defined as the ratio of content words to total tokens:

$$\text{Lex} = \frac{\#\{\text{ADJ, ADV, NOUN, PROPN, VERB}\}}{\#\{\text{all tokens}\}}.$$

We obtain POS tags using Stanza.

Len. Len measures whether the method preserves native-like length variation:

$$\text{Len} = \frac{\sigma(\text{method lengths})}{\sigma(\text{reference lengths})}.$$

Values closer to 1.0 indicate that the method better matches the length variance of human references.

PPL. PPL denotes the mean per-text perplexity of method outputs. While Li et al. (2025) use Llama-3.1-8B, we use Qwen2.5-7B as a multilingual open-weight alternative. Lower PPL is interpreted as more natural generation from the perspective of the language model. (Used in Table 5.)

Stylometric distance to reference. We analyze how much each method output differs from the human reference using BERTScore F1, sentence embedding cosine similarity, LexJSD, and differences in length and punctuation usage. BERTScore F1 measures semantic similarity with the reference using XLM-RoBERTa-Large. Sentence embedding cosine similarity is the cosine similarity between the corpus-level mean embeddings of method outputs and references, obtained using `paraphrase-multilingual-mpnet-base-v2`. LexJSD is the Jensen-Shannon divergence between the token distributions of method outputs and references, where lower values indicate greater similarity to the lexical distribution of the reference. Δlen denotes the relative difference in output length compared to the reference, while Δcomma and Δem-dash represent how much more or less frequently commas and em-dashes are used relative to the reference, measured per 100 characters. Detailed results are provided in Appendix Table 11.

Faithfulness and language-specific calibration. Faithfulness is determined by bidirectional NLI to assess whether meaning is preserved before and after editing. We compute forward and backward entailment probabilities using a multilingual NLI classifier (based on XNLI / KorNLI / OCNLI), and an edit is judged as *faithful* when both values meet or exceed the language-specific threshold $\tau_{\text{nli}}^{(\ell)}$ is selected for each language to satisfy 95%-tile precision using paired human references from the FLORES-200 dev split. Used in Equation 5 and Table 9.

D Translationese Taxonomy in Details

The translationese taxonomy used in this work consists of five top-level categories, namely faithfulness, voice, register, naturalness, and surface form, along with their sub-components. Each category specifies (i) the type of detectable signal, (ii) the measurement method (surface / NLI / LLM judge), and (iii) the evaluation comparison baseline.

Faithfulness. Alignment with the meaning of the source text. Sub-components include semantic entailment (bidirectional NLI), numeric and named-entity accuracy, and LLM judge factual preservation. The primary measurement methods are bidirectional NLI entailment (Maynez et al., 2020) and LLM judge. This serves as a hard constraint for all methods, with mandatory rollback upon failure.

Voice. Preservation of the speaker's idiolect. Sub-components include voice anchor (author anchor consistency, stylometric distance), voice drift (change in idiolect distance before and after editing), and register-anchor consistency. Activated only in Monolingual test set, and measured using stylometric centroid distance combined with native judge voice evaluation.

Register. Consistency of honorifics, keigo, and formal/informal register. Sub-components include intra-document register consistency, formality alignment with source intent, and conformity to the

native register distribution of the target language. The primary measurement methods are phrase-level lexicon (Ko ending classification, Zh 把/被 frequency, etc.) and LLM judge.

Naturalness. Multi-faceted naturalness. Sub-components include lexical naturalness (JSD with native distribution), word order naturalness, discourse marker appropriateness, hedge/emphasis balance (post-training distortion detection), and discourse structure naturalness. In terms of measurement, lexical naturalness is evaluated using LexJSD, hedge/emphasis balance using epistemic-marker frequency, and the remaining sub-components using an LLM judge.

Surface form. Surface-level artifacts such as punctuation, formatting, whitespace, and character normalization are corrected. Sub-components include comma over-use (z-score), dash overuse, English-style colons, forced bullets, and NFC, full-width, half-width, and whitespace normalization, all of which are handled through deterministic surface metrics and a rule-based fixer without affecting meaning.

E Hyperparameters

We summarize the default hyperparameters of MELT in Table 7.

F Model Cards

Table 8 summarizes all models used in this paper according to their roles.

G Per-language Metric Breakdown

In addition to the prefer-edit rate reported in Table 2 of §3.3, the method-level breakdown of NLI entailment rate is provided in Table 9, while the computational cost and efficiency are summarized in Table 10.

From a cost–latency perspective, MELT-LLM and MELT-Hybrid add only one additional model call compared to the single-pass baseline, resulting in an increase in latency of approximately 2 seconds. MELT-MAS is designed for high-accuracy operation and exhibits higher latency due to local inference with a 32B native-judge model, while its API cost remains only USD 1.80 for 1,012 samples. The debate gate is activated only for a subset of samples rather than the entire dataset, and overall, MELT-Hybrid provides a practical cost–performance trade-off.

Stylometric divergence from human reference. To quantify translationese, we measure how much each method's output differs from the FLORES human reference along three dimensions, namely semantic similarity, lexical distribution, and surface markers (Table 11). BERTScore F1 remains within a narrow range of 0.948−0.968 across all methods, indicating that differences between methods arise primarily from surface distributional differences rather than meaning preservation. In terms of length, comma frequency, and lexical distribution, polishing-based methods produce longer outputs and higher comma frequency than the reference across several languages, suggesting that the "Polishing PE > Oracle PE" phenomenon observed in §3.3 may be interpreted as LLM-typical stylistic preference rather than genuine semantic quality improvement. MELT variants show a general tendency to suppress verbosity and comma overuse. LexJSD does not decrease consistently across all languages, and is therefore interpreted as an effect of correcting surface markers directly related to translationese, such as length and punctuation, rather than alignment of the overall lexical distribution.

H Statistical Significance Analysis

Table 12 shows statistical significance analyses for the main results in Table 2. Statistical significance is assessed on the per-sample mean Likert difference, which is the continuous quantity underlying the binary Prefer-Edit Rate (PE). We compute 95% paired bootstrap confidence intervals using 2,000 paired bootstrap resamples and perform paired sign-flip permutation tests. Each entry reports the mean Likert difference (Δ, multiplied by 100) together with its 95% confidence interval.

all MELT variants significantly outperform Self-Refine across the five target languages. Compared with LiL Polish, statistically significant improvements are consistently observed for Ko and Ja, whereas differences in De and Fr are not statistically significant. Zh exhibits mixed behavior across MELT variants, with MELT-LLM significantly outperforming LiL Polish, while MELT-Hybrid and MELT-MAS significantly favor LiL Polish. These results suggest that the apparent preference differences between MELT and LiL Polish in De and Fr fall within the range of sampling variability rather than reflecting consistent performance differences.

Symbol	Stage	Value	Description
τ_z	Detect (surface)	2.0	comma over-use z-threshold
τ_{IoU}	Detect (merge)	0.5	hybrid merge IoU cutoff
τ_{deg}	Verify (surface)	0.5	surface gate degradation tolerance
$\tau_{\text{nli}}^{(\ell)}$	Verify (NLI)	[0.65, 0.75]	bidirectional NLI threshold (per language)
$w_n^{(d)}$	Debate	0.6	native priority weight ($d \in \mathcal{D}_{\text{sel}}$)
τ_{debate}	Debate	0.5	debate gate confidence threshold
δ_c	Debate	0.3	round-1 consensus tolerance

Table 7: Default hyperparameter values for MELT. All values are determined on a 200-sentence pilot subset of FLORES-Ko/Zh and applied directly to the remaining three languages. Only the per-language $\tau_{\text{nli}}^{(\ell)}$ is calibrated per language.

I Calibration Metrics

We report a per-language decomposition of the $\overline{|\Delta_O|}$ (average across the five languages) from Table 2 in §3.3.

The Δ_O of MELT-MAS is most accurately calibrated to Oracle PE across four languages (Ko, Ja, De, Fr) within $|\Delta| \leq 3$, whereas LiL Polish shows inconsistent calibration with over-shoot of $+6.9/+9.3$ in De/Fr and under-shoot in Ja/Zh. In particular, Zh is an outlier with Δ_O values 1.5–2 times larger than other languages (Ko 25.3 vs. Zh 45.8), and all methods under-shoot in Zh due to a Zh-specific reference distribution gap. In terms of LC-PE, LiL Polish shows the highest performance, suggesting that judge preference may be influenced not only by verbosity but also by more general stylistic bias. Table 13 reports the per-language calibration metrics on FLORES.

Table 14 reports the results of an additional human evaluation conducted on 10% of the evaluation data for each target language (101 samples per language) to examine whether the proposed calibration metric is associated with human-perceived naturalness. Native speakers of each target language directly evaluate the naturalness of translations produced by LiL Polish and MELT, with the human reference serving as the Oracle condition. MELT achieves a mean calibration error of 4.0pp in the human evaluation, which is 1.8pp lower than the 5.8pp of LiL Polish. This provides evidence that MELT's lower calibration error is not solely attributable to LLM judge preferences, but also corresponds to a preference distribution closer to that of the human reference under direct human judgments of naturalness.

J Per-rubric Mean Likert Score

Tables 21, 22, and 23 report the mean Likert scores for each rubric assigned by the three LLM judges.

K Judge-Free Stylometric Analysis of Translationese Recovery

To complement LLM-judge-based evaluation, we additionally conduct a judge-free stylometric analysis using classical authorship attribution techniques (Patchala and Bhatnagar, 2018).

K.1 Human-reference Style Recovery in the Multilingual Translation Setting.

We apply the same stylometric probe to the FLORES translation benchmark to analyze whether MELT shifts raw MT toward the human-reference translation style. Since FLORES does not contain author labels, we reformulate the problem as a 2-class style discrimination task that distinguishes human references (H) from raw MT (M), replacing the conventional authorship attribution setup. We construct character trigram, POS trigram, and function-word-based templates from Oracle references and raw MT, and assign the stylometric signature of each output to the nearest template based on KL divergence. We measure the proportion of outputs classified as Oracle-reference style in Table 16.

The experimental results show that only 36.4% of raw MT outputs are classified as human-reference style based on character trigrams, whereas actual human translations record 60.0%, confirming that distinguishable stylometric differences exist between human translations and MT even in the direct MT setting of FLORES. All editing methods shift raw MT toward the human-reference style, with MELT-LLM showing the largest improvement. This suggests that training-

Role	Display name	Provider / full ID
Editor (post-editing generation)		
Editor (all methods)	GPT-4o-mini	`openai/gpt-4o-mini`
LLM judges (9-rubric pairwise, position-swap)		
Default judge	Claude-haiku 4.5	`anthropic/claude-haiku-4-5`
Cross-judge	Gemini-2.5-Pro	`google/gemini-2.5-pro`
Cross-judge	GPT-5	`openai/gpt-5`
Native judges (selective debate, Ko/Zh/Ja/De/Fr)		
Ko native judge	EXAONE-4.0-32B	`LGAI-EXAONE/EXAONE-4.0-32B`
Zh/Ja/De/Fr native judge	Qwen3-32B	`Qwen/Qwen3-32B`
Reference-free metric backbones		
PPL (LiL Lex/Len/PPL)	Qwen2.5-7B	`Qwen/Qwen2.5-7B`
BERTScore + MAUVE	XLM-RoBERTa-large	`xlm-roberta-large`
cos similarity	paraphrase-multilingual-MPNet	`sentence-transformers/paraphrase-multilingual-mpnet-base-v2`
Forward / back MT (monolingual setup)		
Forward / back MT	GPT-4o-mini	`openai/gpt-4o-mini`

Table 8: Model cards. All closed-source API calls are routed through OpenRouter and executed under identical conditions of the prompt using nine rubrics, position swap, and temperature $= 0$.

	NLI entailment rate (↑)				
Method	**Ko**	**Zh**	**Ja**	**De**	**Fr**
MT only	.987	.885	1.00	1.00	1.00
LiL Polish (Li et al., 2025)	.993	.885	.999	.998	.999
MELT-Hybrid	.992	.894	.993	.997	.999
MELT-MAS	.991	.892	.989	.999	.999
Oracle (human reference)	1.00	1.00	1.00	1.00	1.00

Table 9: Auxiliary NLI metric reported alongside PE in Table 2. NLI = bidirectional NLI entailment rate based on multilingual XNLI / KorNLI / OCNLI; threshold $\tau_{\text{nli}}^{(\ell)}$ is calibrated to the per-language 95%-tile. Higher values indicate better meaning preservation.

free editing can partially recover general human translation style in the multilingual MT setting. In particular, the human-reference ceiling itself was as low as $\sim$50% in De and Ko, suggesting that human translations and raw MT are stylometrically relatively close in these languages.

K.2 Judge-free Validation: Authorship Attribution and Recovery Ceiling.

As an additional analysis that excludes dependence on LLM judges, we apply classical stylometric authorship attribution (Patchala and Bhatnagar, 2018). Specifically, we construct per-author relative frequency templates based on character trigram, POS trigram, and function-word features, and attribute test texts to the author with the minimum KL divergence. In the experiments, each author's original texts are split into template and test sets, and attribution is performed among approximately 10 candidate authors (Table 17).

The original texts achieve a macro attribution accuracy of 45.9%, reaching 49.8% when character trigrams are used alone. However, after round-trip MT, performance collapses to 20.2%, approaching chance level ($\sim$10.3%). This analysis also reveals a clear recovery ceiling for training-free editing. All editing methods partially shift surface stylometric statistics toward the human-reference direction, but none recover the author-specific idiolect fingerprints lost during the MT process. In practice, no method including MELT achieves authorship recovery that meaningfully exceeds the MT baseline ($\sim$20%). This is interpreted as editing methods normalizing outputs toward a more general fluent translation style rather than reconstructing the style of a specific author. Sentence-level style cosine similarity shows the same tendency (original: 0.284, MT: 0.225, all methods: $\sim$0.221). Taken together, we confirm that training-free editing can recover surface stylometric characteristics such as comma usage, lexical regularity, and phrasing, but cannot

Method	Calls / sample (↓)	Generated tokens / sample (↓)	Latency / sample (s) (↓)	Cost / 1,012 samples (USD) (↓)
Naive Polish	2.0	224	5.3	0.35
LiL Polish (Li et al., 2025)	2.0	237	5.3	0.36
Self-Refine	6.9	738	18.9	0.86
MELT-LLM	3.0	351	7.1	0.54
MELT-Hybrid	3.0	352	7.4	0.55
MELT-MAS	10.3	2,581	98.5	1.80

Table 10: Computational cost and efficiency across methods, averaged over the five target languages. Calls, generated tokens, and end-to-end latency are reported per sample, while total API cost is reported for 1,012 samples. LLM judge costs are excluded from the reported API cost.

Lang	Method	BERT-F1↑	cos↑	LexJSD↓	Δlen	Δcomma	Δem-dash
Ko	MT only	0.949	0.957	0.3564	+0.018	+0.29	+0.001
	Naive Polish	0.949	0.957	0.3581	+0.005	+0.31	+0.000
	LiL Polish	0.949	0.957	0.3603	+0.017	+0.31	+0.000
	MELT-LLM	0.948	0.957	0.3691	−0.046	+0.22	+0.001
	MELT w/o	0.948	0.957	0.3652	−0.039	+0.14	+0.000
	MELT-MAS	0.948	0.958	0.3661	−0.038	+0.17	+0.001
	Self-Refine	0.948	0.957	0.3623	−0.009	+0.21	+0.000
Zh	MT only	0.951	0.927	0.0227	−0.033	+0.11	+0.001
	Naive Polish	0.952	0.928	0.0223	−0.038	+0.21	+0.001
	LiL Polish	0.951	0.926	0.0230	−0.023	+0.20	+0.001
	MELT-LLM	0.950	0.926	0.0233	−0.126	+0.38	+0.001
	MELT w/o	0.950	0.929	0.0226	−0.080	+0.03	+0.000
	MELT-MAS	0.950	0.928	0.0222	−0.077	+0.05	+0.000
	Self-Refine	0.950	0.927	0.0218	−0.057	+0.26	+0.001
Ja	MT only	0.952	0.892	0.0124	+0.041	+0.14	+0.001
	Naive Polish	0.952	0.892	0.0113	+0.033	+0.14	+0.001
	LiL Polish	0.951	0.890	0.0113	+0.051	+0.11	+0.001
	MELT-LLM	0.949	0.892	0.0147	−0.106	+0.18	+0.001
	MELT w/o	0.949	0.892	0.0149	−0.103	+0.19	+0.001
	MELT-MAS	0.950	0.892	0.0141	−0.096	+0.18	+0.001
	Self-Refine	0.949	0.891	0.0110	−0.015	+0.11	+0.001
De	MT only	0.963	0.962	0.1156	−0.002	+0.18	+0.001
	Naive Polish	0.962	0.961	0.1142	−0.001	+0.12	+0.001
	LiL Polish	0.961	0.960	0.1184	+0.019	+0.11	+0.001
	MELT-LLM	0.960	0.961	0.1207	−0.031	+0.10	+0.001
	MELT w/o	0.960	0.960	0.1188	−0.033	+0.07	+0.001
	MELT-MAS	0.960	0.960	0.1211	−0.034	+0.05	+0.001
	Self-Refine	0.959	0.961	0.1207	−0.022	+0.16	+0.002
Fr	MT only	0.968	0.964	0.0623	+0.001	−0.02	−0.004
	Naive Polish	0.967	0.962	0.0644	−0.001	+0.03	−0.006
	LiL Polish	0.965	0.960	0.0703	+0.016	+0.04	−0.005
	MELT-LLM	0.964	0.959	0.0720	−0.047	+0.00	−0.006
	MELT w/o	0.964	0.958	0.0704	−0.045	−0.05	−0.006
	MELT-MAS	0.965	0.960	0.0692	−0.041	−0.05	−0.006
	Self-Refine	0.965	0.959	0.0676	−0.020	−0.00	−0.004

Table 11: Stylometric divergence of each method's output from the FLORES human reference (n=1,012 per language).

Comparison	Ko	Zh	Ja	De	Fr
MELT-LLM − LiL Polish (Li et al., 2025)	+4.2 [2.0, 6.6]	+4.5 [1.8, 7.4]	+8.6 [5.5, 11.6]	-1.0 [-3.2, 1.1]	+0.5 [-1.6, 2.6]
MELT-Hybrid − LiL Polish (Li et al., 2025)	+2.6 [0.2, 5.0]	-4.6 [-7.2, -2.0]	+9.0 [6.0, 11.8]	-1.9 [-4.3, 0.5]	-2.1 [-4.2, 0.0]
MELT-MAS − LiL Polish (Li et al., 2025)	+3.2 [0.8, 5.5]	-4.1 [-6.6, -1.7]	+10.8 [8.0, 13.4]	-0.5 [-2.9, 1.9]	-0.4 [-2.5, 1.7]
MELT-LLM − Self-Refine (Madaan et al., 2023)	+13.6 [10.5, 16.7]	+15.8 [12.9, 18.9]	+20.2 [16.5, 23.9]	+20.1 [16.8, 23.3]	+26.7 [23.8, 29.8]

Table 12: Pairwise differences in Prefer-Edit Rate (PE) between MELT variants and baselines across languages. Values denote absolute percentage-point differences with 95% bootstrap confidence intervals.

	Δ_O (signed, pp)					LC-PE (length-controlled, %)				
Method	**Ko**	**Zh**	**Ja**	**De**	**Fr**	**Ko**	**Zh**	**Ja**	**De**	**Fr**
MT only	−25.3	−45.8	−44.5	−43.0	−40.9	0.0	0.0	0.0	0.0	0.0
Naive Polish	−1.3	−12.1	−15.4	+2.3	+5.7	21.2	33.2	27.5	45.5	46.3
LiL Polish (Li et al., 2025)	−0.1	−10.4	−12.5	+6.9	+9.3	24.9	36.8	33.4	52.3	52.9
Self-Refine (Madaan et al., 2023)	−5.9	−18.5	−15.2	−12.5	−17.8	13.6	24.4	19.5	28.2	19.3
MELT-LLM (ours, ablation)	+3.9	−5.9	+1.9	+2.8	+6.6	14.6	27.7	20.0	42.1	38.7
MELT-Hybrid	+2.4	−21.3	+1.7	+0.5	+2.4	14.9	18.5	20.4	39.9	34.8
MELT-MAS	+0.9	−22.4	+1.1	+1.2	+3.0	13.1	17.8	21.2	40.4	36.2

Table 13: Per-language calibration metrics on FLORES

Method	Ko	Zh	Ja	De	Fr	mean $\Delta_O \downarrow$
LiL Polish	54.2	62.6	64.6	61.5	62.8	5.8
MELT	57.6	65.0	61.3	66.4	68.1	**4.0**
Oracle	57.1	70.7	68.7	71.4	66.9	—

Table 14: Human naturalness preference rates on 10% of the evaluation data for each target language. The final column reports the mean absolute deviation from the Oracle preference distribution.

restore the fine-grained authorial identity removed by MT.

Dual-track Evaluation Metrics. This dissociation supports the evaluation methodology recommended in this paper. LLM judge PE remains a valid user-facing quality proxy for production deployment, as end users also tend to be LLM-style friendly, but the effects of stylistic preservation and translationese removal must be measured separately using stylometric distance metrics. Since MELT's verification gate maintains the stylometric distribution in the direction of the reference without inflating LLM judge PE, the trade-off along both axes is clearly separated and measured.

L Qualitative MELT post-editing examples

Table 18 (phrase-level; Fr, De) and Table 19 (sentence-level; Zh, Ja, Ko) present representative examples showing how translationese patterns in GPT-4o-mini literal translations across five target languages are corrected through MELT's `detect` → `edit` → `verify` pipeline. Red spans indicate translationese traces identified by MELT's typology-aware detector, while blue spans indicate portions rewritten by MELT into target-native expressions. All examples are preferred over the LLM output by all three judges (Claude Haiku 4.5 / Gemini 2.5 Pro / GPT-5) with an avg.avg. score ≥ 0.56; PPL measured by Qwen2.5-7B also decreases in all cases ($-1.1 \sim -5.1$).

Per-category Detection Statistics. Table 20 presents the aggregated results of translationese detected by Φ_{llm} and Φ_{surf} on FLORES raw MT, broken down by top-level taxonomy category (n=1,012/lang). As these figures represent detector-flagged span counts, they are used for relative cross-lingual trend analysis rather than as absolute counts based on human-annotated gold labels. Naturalness is detected at high frequency in Korean, Chinese, and Japanese. Surface form is relatively prominent in Chinese, showing a large gap compared to other languages. Register is strongly concentrated in Korean, and is rarely detected in Chinese and French. This indicates that English-structure mimicry is a language-universal core characteristic of translationese. Faithfulness and Voice show low overall frequency but are relatively distributed across multiple languages. These results demonstrate that translationese exhibits both universal characteristics and typology-conditioned language-specific biases.

Per-category Post-editing Performance. Figure 4 presents a comparison of MELT's post-editing performance broken down by taxonomy

Lang	Method	MAUVE↑	cos↑	BERT-F1↑	LexJSD↓	Δlen	Δcomma	PE H/G/G5↑
Ko								
	MT (round-trip)	0.010	0.851	0.892	0.5333	−0.084	+0.79	—/—/—
	Naive Polish	0.010	0.850	0.892	0.5336	−0.086	+0.77	5.6/ 5.1/ 3.6
	LiL Polish (Li et al., 2025)	0.012	0.851	0.891	0.5346	−0.073	+0.78	9.2/16.8/ 8.7
	Self-Refine (Madaan et al., 2023)	**0.014**	0.850	0.892	0.5333	−0.084	+0.75	1.5/ 3.6/ 2.0
	MELT-LLM	—	0.848	0.891	0.5367	−0.103	+0.67	6.6/12.8/ 5.6
	MELT-Hybrid	0.013	0.848	0.891	0.5359	−0.102	+0.68	7.1/ 7.7/ 6.1
	MELT-MAS	0.010	0.849	0.891	0.5361	−0.100	+0.67	8.2/ 8.2/ 5.6
Zh (TOC entries filtered; n=135)								
	MT (round-trip)	0.018	0.887	0.905	0.2626	+0.030	−1.08	—/—/—
	Naive Polish	0.027	0.882	0.903	0.2650	−0.002	−1.05	11.9/17.0/14.8
	LiL Polish (Li et al., 2025)	**0.030**	0.883	0.903	0.2658	+0.007	−1.02	25.2/54.8/28.9
	Self-Refine (Madaan et al., 2023)	0.015	0.880	0.902	0.2664	−0.011	−1.10	5.9/14.1/ 8.1
	MELT-LLM	—	0.880	0.902	0.2681	−0.029	−1.02	5.9/10.4/ 8.9
	MELT-Hybrid	0.011	0.883	0.902	0.2654	−0.019	−1.07	2.2/ 9.6/ 5.9
	MELT-MAS	0.011	0.883	0.902	0.2655	−0.018	−1.09	5.2/ 6.7/ 5.9
Ja								
	MT (round-trip)	**0.032**	0.834	0.879	0.0608	+0.010	+0.06	—/—/—
	Naive Polish	0.020	0.834	0.878	0.0609	+0.007	+0.07	8.5/11.5/13.0
	LiL Polish (Li et al., 2025)	0.015	0.834	0.878	0.0605	+0.020	+0.06	10.5/43.5/22.0
	Self-Refine (Madaan et al., 2023)	0.020	0.829	0.877	0.0620	−0.017	−0.04	3.5/22.0/13.0
	MELT-LLM	—	0.826	0.877	0.0609	−0.089	−0.04	15.5/36.0/20.5
	MELT-Hybrid	0.023	0.829	0.877	0.0609	−0.092	+0.01	19.0/43.5/23.0
	MELT-MAS	0.018	0.827	0.877	0.0608	−0.092	+0.01	18.0/44.0/24.0
De								
	MT (round-trip)	0.950	0.964	0.963	0.1456	+0.027	−0.04	—/—/—
	Naive Polish	0.893	0.961	0.960	0.1487	+0.029	+0.04	31.0/30.5/28.0
	LiL Polish (Li et al., 2025)	**0.967**	0.962	0.960	0.1525	+0.034	+0.05	37.5/44.5/29.5
	Self-Refine (Madaan et al., 2023)	0.933	0.958	0.958	0.1516	+0.029	−0.00	12.0/10.5/13.0
	MELT-LLM	—	0.959	0.957	0.1518	+0.020	+0.02	20.5/20.5/16.0
	MELT-Hybrid	0.961	0.959	0.958	0.1507	+0.019	−0.04	16.5/21.0/19.0
	MELT-MAS	0.939	0.961	0.958	0.1506	+0.022	−0.02	17.0/17.0/13.5
Fr								
	MT (round-trip)	**0.297**	0.941	0.939	0.3561	+0.020	−0.07	—/—/—
	Naive Polish	0.252	0.934	0.930	0.3661	+0.019	−0.09	28.1/34.1/34.6
	LiL Polish (Li et al., 2025)	0.290	0.931	0.928	0.3707	+0.038	−0.07	29.5/41.9/38.7
	Self-Refine (Madaan et al., 2023)	0.266	0.935	0.934	0.3641	+0.009	−0.09	14.7/13.3/16.5
	MELT-LLM	—	0.931	0.931	0.3710	−0.017	−0.09	23.9/29.4/28.0
	MELT-Hybrid	0.264	0.932	0.931	0.3707	−0.015	−0.13	25.2/30.3/28.0
	MELT-MAS	0.286	0.933	0.932	0.3704	−0.012	−0.12	29.4/36.2/31.2

Table 15: Author-style recovery experiment (Frame A + B combined): distributional alignment and LLM judge preference comparison between the original texts of five languages × 49 pre-LLM-era authors / 972 paragraphs and each method's output. Pipeline: original → English (forward MT) → target language (back MT, "MT round-trip" baseline; reference with strongly injected LLM-ese) → "editing by each method to attempt recovery of the original author's style" PE H/G/G5 = 3-judge prefer-method rate (%, method output vs. round-trip baseline).

Condition	Reference-style rate (% H, 5-lang macro-avg)			
	char3	POS-tri	func	comb.
Human reference (H)	**60.0**	**53.3**	**55.0**	**58.0**
raw MT (M)	36.4	43.9	49.6	40.6
Naive Polish	39.5	43.6	50.6	42.3
LiL polish (Li et al., 2025)	39.8	45.0	50.9	43.0
Self-Refine (Madaan et al., 2023)	41.1	44.9	48.7	42.9
MELT-LLM	**42.8**	45.3	**51.2**	**45.0**
MELT-Hybrid	41.9	46.2	50.5	44.4
MELT-MAS	41.9	**46.5**	50.8	44.9
Chance (2-class)			50.0	

Table 16: Judge-free FLORES style discrimination: 2-class nearest-template classification (human reference H vs raw MT M) over the same features as Table 17 (Patchala and Bhatnagar, 2018). Metric is the reference-style rate (% classified as H), 5-lang macro-average. All editors shift raw MT toward human-reference style and MELT-LLM shifts it most.

Condition	Attribution acc. (%, 5-lang macro-avg)			
	char3	POS-tri	func	comb.
Original (author)	**49.8**	**28.7**	**22.6**	**45.9**
MT round-trip	22.2	16.2	16.9	20.2
Naive Polish	22.3	16.3	16.9	19.4
LiL polish (Li et al., 2025)	20.9	17.4	15.6	20.6
Self-Refine (Madaan et al., 2023)	22.9	16.5	16.6	20.6
MELT-LLM	22.2	16.7	15.9	21.4
MELT-Hybrid	22.4	16.6	15.5	20.6
MELT-MAS	22.4	18.2	14.7	20.2
Chance			≈ 10.3	

Table 17: Judge-free authorship attribution accuracy (Patchala and Bhatnagar, 2018; per-author KL template matching over character trigrams / POS trigrams / function words / combined), 5-lang macro-average over $\sim$10 candidate authors. MT Round-trip collapses author identifiability toward chance and no training-free editor restores it above the MT floor.

category for translationese detected by Φ_{llm} and Φ_{surf} (five languages pooled, 3-judge average PE). All MELT variants show consistent improvements over raw MT across all categories. In particular, MELT-LLM records high PE in most categories, with notably large improvements in naturalness and surface form. This is consistent with the calibration results in Table 2.

M Cross-family Editor Evaluation

Table 24 shows the results across three editor families. GPT-4o-mini, Claude Haiku 4.5, and Gemini 2.5 Flash are used as editors while keeping the remaining MELT pipeline unchanged. All methods are evaluated using a single Claude Haiku 4.5 judge.

MELT exhibits stable performance across different editor families. MELT-LLM or MELT-Hybrid achieves higher PE than LiL Polish in 11 of the 15 editor-language combinations. The MAUVE metric also remains consistently high across all editor families, indicating stable distributional alignment with human references regardless of the editor model. In contrast, LiL Polish shows substantial degradation under the Claude Haiku 4.5 editor, where MAUVE drops to 0.03 (Zh) and 0.31 (Ja). Inspection of the generated outputs reveals that LiL Polish frequently returns the original MT without editing under this editor (no-edit rates of 92%, 82%, and 57% for Ko, De, and Ja, respectively), suggesting that its behavior is more sensitive to the prompt–editor combination. Calibration results show that MELT achieves the smallest $\overline{|\Delta_O|}$ under GPT-4o-mini and Claude Haiku 4.5, whereas all methods overshoot the Oracle under Gemini 2.5 Flash. This suggests that the calibration behavior is more dependent on the editor family than the judge-free distributional alignment measured by MAUVE.

Phrase-level LLM-ese: structural calque on a single clause	
Source	Travellers are strongly advised to be aware of any risk of severe weather affecting their area...
LLM	Les voyageurs sont fortement conseillés de se renseigner sur tout risque de conditions météorologiques sévères affectant leur région... (PPL: **7.4**)
MELT	Il est recommandé aux voyageurs de se tenir informés des risques de conditions météorologiques sévères dans leur région... (PPL: **5.6** ↓)
DETECTOR: `B1_voice_shift`: French does not license a person-subject *être conseillé*; MELT restores the impersonal *Il est recommandé à*.	*Fr-877*
Source	No immediate reports of damage have been received by the United States Geological Survey (USGS) and its National Earthquake Information Center.
LLM	Es liegen keine unmittelbaren Berichte über Schäden vor, die vom United States Geological Survey (USGS) und seinem National Earthquake Information Center erhalten wurden. (PPL: **14.4**)
MELT	Es wurden keine unmittelbaren Berichte über Schäden vom Geologischen Dienst der Vereinigten Staaten und seinem Nationalen Erdbebeninformationszentrum erhalten. (PPL: **9.3** ↓)
DETECTOR: `D1_phrasing_unnatural`: English passive "*reports have been received by*" calqued; English institution names left untranslated. MELT inverts to the German active-passive "*Es wurden... erhalten*" and translates the proper nouns.	*De-182*

Table 18: Phrase-level MELT post-editing examples. Single-clause anglicisms (red) flagged by MELT's typology-aware detector and rewritten into target-native phrasing. PPL drops with the rewrite. See Table 19 for sentence-level cases.

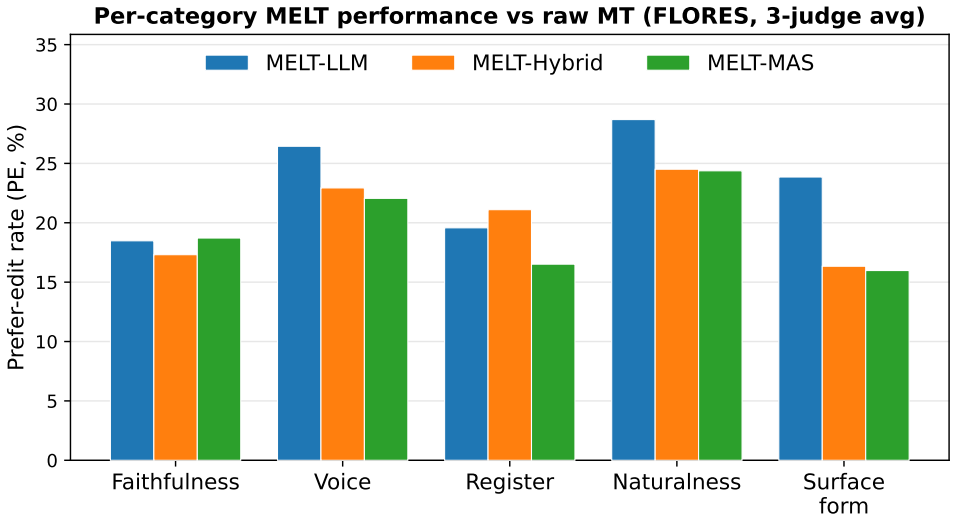

Figure 4: Per-category MELT performance on detector-flagged subsets from FLORES raw MT (pooled across the five languages). Bars denote PE averaged across three judges (Claude Haiku 4.5, Gemini-Pro, GPT-5). All MELT variants consistently improve over raw MT, while MELT-LLM achieves the strongest gains in most categories, particularly naturalness and surface form.

Sentence-level LLM-ese: clause-coordination and register mismatch

Source As NSA, he assisted Carter in diplomatically handling world affairs, such as the Camp David Accords, 1978; normalizing US-China relations...

LLM 作家安全，他助特在外交上理世界事，例如1978年的戴；在1970年代末期正常化美中系；伊朗革命，致1979年的伊朗人危机；... (PPL: **10.1**)

MELT 他任家安全期，助特理多事，包括1978年的戴 1970年代末美中系正常化 1979年的伊朗革命及其引的伊朗人危机... (PPL: **5.5** ↓)

DETECTOR: `D1_phrasing_unnatural` + `D2_connective_overuse`: light-verb "作" + semicolon-separated English-style clauses + over-formal "致". MELT replaces with native "任...期" and merges the list with " " / "及其引的". *Zh-019*

Source Mass car ownership also leads to a higher incidence of accidents on the roads, which leads to the invention of new techniques in healthcare...

LLM 大量の自動車所有は 道路上での事故の生率を高め それが損傷した身を修復するための新しい療技術の明につながります 。(PPL: **9.5**)

MELT 自動車の普及は道路での事故を加させ 損傷した身を修復する新しい療技術の明につながる 。(PPL: **8.4** ↓)

DETECTOR: `D1_phrasing_unnatural` + `D2_connective_overuse`: literal "大量の自動車所有" + redundant pronoun-bridge "それが" calques English "*which leads to*". MELT uses native "自動車の普及" + verb-chain coordination, removing the bridge entirely. *Ja-485*

Source Blogs can also help improve student writing. While students often begin their blog experience with sloppy grammar and spelling, the presence of an audience generally changes that.

LLM 블로그는 학생들의 글쓰기 향상에도 도움이 될 수 있습니다. 학생들이 종종 엉성한 문법과 철자로 블로그 경험을 시작하지만, 청중이 존재하면 일반적으로 그런 점이 변화합니다. (PPL: **10.1**)

MELT 블로그는 학생들의 글쓰기를 향상시키는 데 도움이 됩니다. 학생들은 종종 엉성한 문법과 철자로 블로그를 시작하지만, 청중이 있으면 그런 점이 달라집니다. (PPL: **5.9** ↓)

DETECTOR: `C1_honorific_mixing` + `D1_phrasing_unnatural`: register over-formal for encyclopedic content + literal "블로그 경험을 시작" + verbose "존재하면 일반적으로...변화합니다". MELT compresses to "있으면...달라집니다". *Ko-687*

Table 19: Sentence-level MELT post-editing examples. Multi-clause LLM-ese involving coordination calque (Zh/Ja) or register/encyclopedic-tone mismatch (Ko). Every example is preferred over the LLM baseline by all three judges with avg. score ≥ 0.56; PPL drops accompany every rewrite.

Category	Ko	Zh	Ja	De	Fr	All
Faithfulness	49	59	75	52	62	297
Voice	84	61	64	42	24	275
Register	**177**	0	13	2	0	192
Naturalness	1563	1630	1354	927	518	5992
Surface form	333	**828**	119	165	126	1571
Total	2206	2578	1625	1188	730	8327

Table 20: LLM-ese category distribution in raw MT (FLORES devtest, n=1012/lang). Aggregated *detector-flagged span counts* detected by Φ_{llm} and Φ_{surf} (§2.2) in raw GPT-4o-mini translations y_{mt}, broken down by top-level taxonomy category. These are automatically detected results, not human-annotated gold labels.

Lang	Method	Faith.	Adeq.	Flu.	Syn.	Lex.	Disc.	Reg.	Conc.	Human	Avg
Ko	MT only	4.78	4.80	4.39	4.16	4.21	4.11	4.48	3.72	3.89	4.28
	LiL Polish	4.77	4.77	4.68	4.60	4.44	4.53	4.70	4.38	4.37	4.58
	Self-Refine	4.57	4.58	4.35	4.24	4.30	4.27	4.44	4.19	4.10	4.34
	MELT-Hybrid	4.58	4.57	4.59	4.49	4.47	4.48	4.61	4.56	4.41	4.53
	MELT-MAS	4.60	4.58	4.61	4.52	4.50	4.52	4.63	4.59	4.46	4.56
	Oracle (human ref)	4.41	4.45	3.98	3.73	3.97	3.86	4.08	3.45	3.39	3.92
Zh	MT only	4.76	4.83	4.48	4.15	4.18	4.33	4.73	3.69	3.91	4.34
	LiL Polish	4.79	4.85	4.79	4.68	4.51	4.74	4.89	4.42	4.46	4.68
	Self-Refine	4.56	4.67	4.53	4.40	4.23	4.53	4.78	4.26	4.21	4.46
	MELT-Hybrid	4.54	4.59	4.74	4.70	4.54	4.66	4.91	4.73	4.54	4.66
	MELT-MAS	4.55	4.60	4.75	4.71	4.57	4.67	4.92	4.75	4.58	4.68
	Oracle (human ref)	4.72	4.80	4.67	4.42	4.40	4.53	4.90	3.81	4.23	4.50
Ja	MT only	4.75	4.76	4.23	3.93	3.81	3.87	3.74	3.29	3.48	3.98
	LiL Polish	4.72	4.73	4.63	4.50	4.29	4.42	4.53	4.13	4.28	4.47
	Self-Refine	4.39	4.38	4.24	4.13	4.01	4.08	4.19	3.99	3.95	4.15
	MELT-Hybrid	4.47	4.45	4.50	4.49	4.38	4.31	4.27	4.66	4.33	4.43
	MELT-MAS	4.47	4.44	4.52	4.50	4.40	4.34	4.31	4.65	4.40	4.45
	Oracle (human ref)	4.58	4.59	4.38	4.23	4.06	3.96	3.81	3.47	3.81	4.10
De	MT only	4.69	4.83	4.53	4.29	4.21	4.44	4.74	3.87	3.88	4.39
	LiL Polish	4.79	4.89	4.89	4.80	4.69	4.82	4.89	4.48	4.62	4.76
	Self-Refine	4.48	4.66	4.64	4.47	4.37	4.63	4.75	4.24	4.27	4.50
	MELT-Hybrid	4.62	4.74	4.85	4.73	4.62	4.75	4.82	4.54	4.59	4.70
	MELT-MAS	4.64	4.76	4.86	4.76	4.67	4.77	4.83	4.59	4.66	4.73
	Oracle (human ref)	4.62	4.77	4.64	4.53	4.45	4.54	4.79	4.01	4.13	4.50
Fr	MT only	4.75	4.92	4.72	4.27	4.28	4.36	4.81	4.06	4.06	4.47
	LiL Polish	4.68	4.90	4.94	4.84	4.66	4.76	4.95	4.57	4.66	4.77
	Self-Refine	4.42	4.69	4.73	4.48	4.30	4.49	4.85	4.42	4.26	4.52
	MELT-Hybrid	4.48	4.69	4.93	4.84	4.67	4.64	4.91	4.77	4.67	4.73
	MELT-MAS	4.52	4.72	4.96	4.89	4.72	4.70	4.92	4.79	4.75	4.77
	Oracle (human ref)	4.53	4.77	4.79	4.47	4.35	4.39	4.85	4.03	4.22	4.49

Table 21: Per-rubric mean Likert score (1–5, ↑) assigned by **Claude Haiku 4.5** using nine rubrics. Each cell represents the mean 1–5 Likert score assigned by the LLM judge to the output of the corresponding method (Oracle = human reference). The final Avg column is the simple average across the nine rubrics. Rubric definitions: Faith.=faithfulness, Adeq.=adequacy, Flu.=fluency, Syn.=syntax, Lex.=lexicon, Disc.=discourse, Reg.=register/punctuation, Conc.=conciseness, Human=human-likeness. For the LLM judge prompt definition using nine rubrics.Table 2 is computed based on the Likert scores in this table, while this table reports the mean Likert scores themselves for each method.

Lang	Method	Faith.	Adeq.	Flu.	Syn.	Lex.	Disc.	Reg.	Conc.	Human	Avg
Ko	MT only	4.66	4.81	4.42	4.31	4.41	4.82	4.92	4.40	4.26	4.56
	LiL Polish	4.80	4.92	4.80	4.82	4.79	4.96	4.99	4.75	4.68	4.83
	Self-Refine	4.53	4.69	4.59	4.66	4.69	4.90	4.94	4.64	4.46	4.68
	MELT-Hybrid	4.57	4.68	4.79	4.84	4.84	4.95	4.99	4.91	4.67	4.80
	MELT-MAS	4.61	4.68	4.82	4.85	4.84	4.94	4.99	4.92	4.72	4.82
	Oracle (human ref)	4.23	4.48	4.33	4.41	4.30	4.86	4.78	4.42	4.18	4.44
Zh	MT only	4.61	4.81	4.52	4.29	4.31	4.78	4.91	4.40	4.19	4.54
	LiL Polish	4.77	4.92	4.90	4.81	4.74	4.98	4.99	4.68	4.74	4.84
	Self-Refine	4.60	4.78	4.73	4.68	4.56	4.96	4.98	4.63	4.53	4.72
	MELT-Hybrid	4.53	4.65	4.94	4.93	4.71	4.92	4.99	4.91	4.78	4.82
	MELT-MAS	4.56	4.67	4.94	4.89	4.74	4.94	5.00	4.92	4.79	4.83
	Oracle (human ref)	4.54	4.77	4.80	4.71	4.56	4.94	4.92	4.54	4.61	4.71
Ja	MT only	4.67	4.80	4.20	4.22	4.17	4.88	4.92	4.26	4.05	4.46
	LiL Polish	4.81	4.92	4.79	4.89	4.76	4.97	4.98	4.70	4.70	4.84
	Self-Refine	4.48	4.62	4.53	4.72	4.62	4.92	4.91	4.61	4.44	4.65
	MELT-Hybrid	4.49	4.55	4.82	4.94	4.88	4.96	4.92	4.95	4.76	4.81
	MELT-MAS	4.50	4.56	4.84	4.94	4.89	4.96	4.94	4.96	4.79	4.82
	Oracle (human ref)	4.37	4.53	4.63	4.77	4.58	4.94	4.93	4.55	4.53	4.65
De	MT only	4.66	4.91	4.41	4.55	4.28	4.96	4.86	4.52	4.21	4.60
	LiL Polish	4.80	4.98	4.92	4.97	4.81	4.99	4.95	4.78	4.84	4.89
	Self-Refine	4.52	4.79	4.60	4.75	4.52	4.99	4.93	4.64	4.48	4.69
	MELT-Hybrid	4.64	4.87	4.90	4.95	4.77	4.99	4.96	4.85	4.81	4.86
	MELT-MAS	4.66	4.87	4.91	4.96	4.79	4.99	4.96	4.87	4.84	4.87
	Oracle (human ref)	4.57	4.87	4.68	4.83	4.55	4.98	4.93	4.56	4.56	4.73
Fr	MT only	4.64	4.89	4.60	4.58	4.41	4.84	4.86	4.69	4.28	4.64
	LiL Polish	4.73	4.96	4.96	4.96	4.81	4.99	4.95	4.74	4.87	4.89
	Self-Refine	4.43	4.79	4.66	4.74	4.49	4.96	4.93	4.66	4.43	4.68
	MELT-Hybrid	4.51	4.79	4.95	4.95	4.78	4.97	4.99	4.91	4.83	4.85
	MELT-MAS	4.56	4.79	4.97	4.98	4.82	4.97	4.99	4.91	4.88	4.87
	Oracle (human ref)	4.42	4.80	4.79	4.83	4.52	4.94	4.95	4.60	4.58	4.71

Table 22: Per-rubric mean Likert score (1–5, ↑) assigned by **Gemini 2.5 Pro** using nine rubrics. All other settings are identical to Table 21.

Lang	Method	Faith.	Adeq.	Flu.	Syn.	Lex.	Disc.	Reg.	Conc.	Human	Avg
Ko	MT only	4.49	4.72	4.45	4.39	4.39	4.68	4.67	4.29	4.41	4.50
	LiL Polish	4.61	4.82	4.74	4.67	4.67	4.91	4.93	4.53	4.71	4.73
	Self-Refine	4.43	4.65	4.56	4.52	4.54	4.80	4.83	4.50	4.54	4.60
	MELT-Hybrid	4.46	4.65	4.74	4.70	4.74	4.90	4.93	4.77	4.74	4.74
	MELT-MAS	4.48	4.64	4.77	4.73	4.76	4.90	4.93	4.79	4.76	4.75
	Oracle (human ref)	3.90	4.23	4.01	4.00	3.97	4.40	4.30	3.98	3.99	4.09
Zh	MT only	4.44	4.66	4.57	4.53	4.38	4.64	4.70	4.47	4.48	4.54
	LiL Polish	4.58	4.78	4.86	4.82	4.68	4.91	4.96	4.67	4.83	4.79
	Self-Refine	4.36	4.59	4.70	4.65	4.44	4.79	4.89	4.63	4.65	4.63
	MELT-Hybrid	4.36	4.51	4.88	4.85	4.69	4.89	4.97	4.92	4.86	4.77
	MELT-MAS	4.40	4.54	4.88	4.85	4.73	4.90	4.97	4.91	4.86	4.78
	Oracle (human ref)	4.21	4.46	4.60	4.58	4.31	4.67	4.68	4.43	4.54	4.50
Ja	MT only	4.50	4.69	4.43	4.44	4.25	4.66	4.82	4.24	4.37	4.49
	LiL Polish	4.72	4.85	4.80	4.81	4.65	4.91	4.96	4.52	4.76	4.78
	Self-Refine	4.38	4.53	4.53	4.54	4.39	4.71	4.83	4.50	4.51	4.55
	MELT-Hybrid	4.40	4.51	4.68	4.71	4.67	4.83	4.82	4.90	4.68	4.69
	MELT-MAS	4.45	4.55	4.69	4.72	4.70	4.84	4.84	4.89	4.70	4.71
	Oracle (human ref)	4.17	4.34	4.44	4.45	4.22	4.61	4.76	4.22	4.40	4.40
De	MT only	4.49	4.80	4.48	4.62	4.27	4.84	4.78	4.39	4.40	4.56
	LiL Polish	4.70	4.91	4.88	4.94	4.73	4.98	4.91	4.68	4.85	4.84
	Self-Refine	4.34	4.67	4.60	4.72	4.37	4.91	4.85	4.54	4.55	4.62
	MELT-Hybrid	4.56	4.80	4.86	4.92	4.69	4.98	4.92	4.78	4.83	4.82
	MELT-MAS	4.57	4.81	4.87	4.92	4.72	4.98	4.90	4.81	4.85	4.83
	Oracle (human ref)	4.38	4.73	4.59	4.75	4.36	4.89	4.81	4.46	4.54	4.61
Fr	MT only	4.53	4.90	4.56	4.50	4.37	4.69	4.72	4.60	4.50	4.60
	LiL Polish	4.71	4.95	4.92	4.91	4.79	4.97	4.93	4.78	4.91	4.87
	Self-Refine	4.40	4.81	4.64	4.60	4.40	4.84	4.83	4.67	4.61	4.64
	MELT-Hybrid	4.58	4.86	4.91	4.89	4.78	4.96	4.97	4.91	4.90	4.86
	MELT-MAS	4.61	4.87	4.93	4.92	4.80	4.96	4.98	4.91	4.92	4.88
	Oracle (human ref)	4.33	4.75	4.67	4.65	4.43	4.79	4.84	4.60	4.65	4.63

Table 23: Per-rubric mean Likert score (1–5, ↑) assigned by **GPT-5** using nine rubrics. All other settings are identical to Table 21.

Method	Ko		Zh		Ja		De		Fr		
	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	PE↑	MAUVE↑	$\overline{\lvert\Delta_O\rvert}$↓
GPT-4o-mini											
LiL Polish (Li et al., 2025)	25.3	.90	35.4	.90	31.9	.93	49.9	.97	50.2	.98	7.8
MELT-LLM	29.2	.94	39.8	.90	46.3	.90	45.8	.95	47.4	.97	**4.2**
MELT-Hybrid	27.8	.91	24.4	.93	46.1	.93	43.5	.96	43.3	.98	5.7
Claude Haiku 4.5											
LiL Polish (Li et al., 2025)	5.8	.81	65.5	**.03**	29.6	**.31**	11.6	.72	11.9	.79	22.9
MELT-LLM	39.1	.93	53.1	.90	66.4	.93	55.3	.96	48.9	.97	12.7
MELT-Hybrid	34.4	.93	32.5	.94	65.5	.92	53.8	.96	45.7	.97	**11.8**
Gemini 2.5 Flash											
LiL Polish (Li et al., 2025)	30.3	.89	63.3	.72	57.9	.90	53.0	.94	50.3	.93	**11.1**
MELT-LLM	45.4	.94	62.6	.92	73.7	.95	65.8	.97	63.6	.97	22.3
MELT-Hybrid	39.7	.92	30.2	.92	73.7	.94	62.8	.98	53.5	.97	18.3
Oracle (Haiku 4.5)	25.3	—	45.8	—	44.5	—	43.0	—	40.9	—	—

Table 24: Editor generalization results across three LLM families.