KUEST: Korean Game Usage Evaluation Suite for LLM Translation

Minjoo Choi* **Jaejun Jung*** **Gunwoo Han*** **Junhyeop Kim***
Yiji Yoon* **Hyemi Kim*** **Yeonseo Cho*** **Jinwoo Jang†**

Pseudo Lab

mjchoi0831@gmail.com anayana9988@gmail.com aya02245@gmail.com junheab12@gmail.com
chokogangdo@gmail.com hyemikim0720@gmail.com samcholab@gmail.com wdfrty1234@gmail.com

Abstract

Despite the growing demand for multilingual game localization, systematic evaluation frameworks for game translation constraints remain largely unexplored. As part of the WMT26 Test Suite Track, we propose **KUEST**, a test suite comprising 1,796 English-to-Korean segments across four core categories, and evaluate 27 submitted systems. Experimental results demonstrate that standard metrics obscure true model capabilities (e.g., the overall 1st-place system dropping to 15th in creative translation). Furthermore, we reveal residual capability fragmentation—category-specific variance that survives the shared general-competence factor and inverts top-tier rankings—and observe a "prompt gap" where context metadata degrades performance. As such, KUEST serves not only as a test suite but also as an evaluation framework that multi-dimensionally dissects domain-specific translation blind spots hidden behind single composite scores. The dataset is available at JudyChoi/KUEST[1].

1 Introduction

Recent advancements in large language models (LLMs) have transformed the machine translation paradigm (Kocmi et al., 2024), a trend particularly prominent in the video game industry, which faces tight schedules and massive multilingual demands (Bernal-Merino, 2015; Hansen and Houlmont, 2022).

However, general-purpose translation and video game translation require that we "mind the gap" between their distinct domain requirements. Game localization entails multi-faceted constraints, including: (1) strict compliance with canonical glossaries (Post and Vilar, 2018; Dinu et al., 2019), (2) precise reproduction of colloquial nuances reflecting speaker personas and interaction contexts (Shterionov et al., 2020), (3) robustness under technical constraints such as UI length limits and markup tag preservation (Specia et al., 2021), and (4) cultural transcreation tailored for resonance and appropriateness (Mangiron, 2016).

These unique characteristics create critical "blind spots" uncaptured by conventional translation evaluation methodologies. To bridge this gap, we propose **KUEST** (**K**orean game **U**sage **E**valuation **S**uite for LLM **T**ranslation) as part of the WMT26 Test Suite Track. KUEST comprises 1,796 English–Korean (En–Ko) game source segments, evaluating system performance across four primary axes:

- **Terminology & Accuracy (TA) (331 segments):** Canonical glossary compliance and word sense disambiguation.
- **Persona & Pragmatics (PP) (388 segments):** Consistency in speech styles and honorifics based on personas and relationships.
- **Technical Robustness (TR) (507 segments):** Placeholder/tag preservation and stability under UI length constraints.
- **Cultural Transcreation (CT) (570 segments):** Cultural adaptation capabilities, including idioms, contextual rendering of hate speech, and gender-neutral translation.

This study proposes an evaluation framework utilizing KUEST to assess 27 system outputs provided by WMT according to the specific characteristics of each category.

2 Related Work

In this section, we review prior literature on game translation, phenomenon-based MT evaluation, KUEST's evaluation domains, and automatic LLM judgment.

[1] https://huggingface.co/datasets/JudyChoi/KUEST

Proceedings of the Eleventh Conference on Machine Translation, pages 1484–1505
October 28-29, 2026 ©2026 Association for Computational Linguistics

2.1 Background and Evaluation Methods in Game Translation

Game translation simultaneously demands adherence to technical constraints and creative rewriting across UI, system messages, and narrative (Mangiron and O'Hagan, 2006; Bernal-Merino, 2015). However, prior studies faced challenges in operationalizing these requirements into evaluation criteria, while surface-level metrics and LLM-as-a-Judge approaches fail to deliver consistent evaluations (Moghe et al., 2025).

While reference-free LLM evaluation methods such as GEMBA (Kocmi and Federmann, 2023) have been proposed, KUEST adopts a hybrid evaluation paradigm—combining rule-based verification with LLM-as-a-Judge—to overcome the limitations of single-score reductions. Specifically, it applies rule-based metrics to formal constraints and category-specific rubrics to semantic judgments.

2.2 Phenomenon-Based Test Suites and KUEST

Test suites isolate fine-grained translation errors that are difficult to diagnose via corpus-level average scores (King and Falkedal, 1990; Isabelle et al., 2017; Macketanz et al., 2018). KUEST adopts this approach to evaluate system capabilities across four core domains:

- **Terminology & Accuracy (TA):** Evaluates contextual Word Sense Disambiguation (WSD) (Raganato et al., 2019) and canonical glossary integration under constrained decoding (Post and Vilar, 2018; Dinu et al., 2019).
- **Persona & Pragmatics (PP):** Assesses Korean honorific consistency and sentence-ending styles based on speaker–listener relationships and discourse contexts (Sohn, 1999; Hwang et al., 2021; Nădejde et al., 2022).
- **Technical Robustness (TR):** Evaluates markup tag preservation (Hanneman and Dinu, 2020), UI length constraint management, and complex structural constraints (Lakew et al., 2019; Zhang et al., 2023).
- **Cultural Transcreation (CT):** Assesses creative rewriting of proper nouns/wordplay (Mangiron and O'Hagan, 2006) and sensitive content adaptation, including gender bias mitigation (Savoldi et al., 2021) and toxicity contextualization (Costa-jussà et al., 2023).

While phenomenon-based evaluation is crucial for game translation, standardized test suites remain lacking. To bridge this gap, we present KUEST, a domain-specific test suite evaluated via a hybrid rule-LLM approach across four target domains (Section 3).

3 Test Suite Creation

This section describes the construction process of KUEST, designed to evaluate the multi-faceted challenges of game translation.

3.1 Raw Data Preprocessing

Game texts generally fall into two primary categories: functional system texts and narrative dialogue. Accordingly, we select two open-source games that represent each axis and align with WMT26 licensing guidelines as our source data:

- **CDDA**[2]: Focuses on functional content, including extensive item descriptions, crafting recipes, and UI elements (average segment length: 38.0 English characters). Approximately 8.0% of segments contain format specifiers (e.g., `%s`, `%d`), making CDDA a primary source for Technical Robustness (TR) segments.
- **Wesnoth**[3]: Focuses on NPC dialogue-driven narrative texts, incorporating 32 mainline text domains with longer segments (average length: 60.8 English characters). Rich speaker metadata (e.g., `[message] speaker=...`) and distinct fantasy tones render it ideal for Persona & Pragmatics (PP) and Cultural Transcreation (CT) segments.

To eliminate reference noise and ensure evaluation reliability, the raw data underwent a two-stage refinement process: (1) removal of untranslated segments (i.e., source strings without corresponding target texts), and (2) removal of fuzzy-flagged segments, which represent tentative translations generated by automatic matching rather than confirmed ones. Both types were fully discarded from the evaluation dataset.

[2]Cataclysm: Dark Days Ahead: https://github.com/harakka/Cataclysm-DDA/blob/6d7c916b46516779c1d5c3ecfc8840ba7d8c108d/lang/po/ko.po

[3]The Battle for Wesnoth: https://github.com/wesnoth/wesnoth/tree/master/po

Table 1: Preprocessing statistics of raw data

Filter Stage / Metric	CDDA	Wesnoth
Total Segments (Non-obsolete)	11,239	26,310
(1) Removed Untranslated	-684	-18,706
(2) Removed Fuzzy Flags	-1,869	-2,088
Final Cleaned 1:1 Pairs	**8,686**	**5,516**
Avg. Length (En / Ko)	38.0 / 21.5	60.8 / 36.3
Format Codes (%s / %d)	8.0%	0.1%
Text Type	UI Item labels	Dialogue Narrative

3.2 Evaluation Categories

3.2.1 Terminology & Accuracy (TA)

Evaluates adherence to domain-specific glossaries and the disambiguation of context-dependent terms (331 segments in total).

- **Glossary Compliance (264 segments):** Evaluates domain glossary adherence, constructed through a 3-stage pipeline of candidate mining, statistical canonical determination, and human inspection.
- **Polysemy (67 segments):** Verifies contextual word sense disambiguation using 19 polysemous words whose meanings shift according to context.

3.2.2 Persona & Pragmatics (PP)

Evaluates Korean speech style consistency (sentence-final honorifics and formality) across speaker role tiers (e.g., Royalty, Military, Commoner) and relational contexts (388 segments).

- **Dual-Prompt Structure:** Evaluates models under a metadata-free *Baseline* and a *Regime R* with factual relational info. Without explicit style directives, it measures autonomous speech style selection across 97 dialogue pairs (2 prompt regimes $\times$ 2 directions = 388 segments).

3.2.3 Technical Robustness (TR)

Evaluates the model's robustness and adherence to technical constraints required by game engines and UI frameworks (507 segments in total).

- **Placeholder Reordering (362 segments):** Measures adaptability to syntactic variations of formatting codes (%s, %d) and variables, split into Consistency (172 segments) and Diverse (190 segments) sets.
- **Tag Preservation (95 segments):** Validates the structural integrity of HTML/XML markup tags.
- **Length Constraint (50 segments):** Evaluates text overflow control using actual font-rendered pixel widths rather than simple character counts, explicitly filtering for cases where Korean rendering width exceeds English ($ko_width > en_width$).

3.2.4 Cultural Transcreation (CT)

Evaluates cultural adaptation capabilities tailored to Korean cultural contexts and gaming sentiments, moving beyond literal translation (8 subcategories, 570 segments in total).

- **Speech Style Adaptation (273 segments):** Evaluates macro-styles (Archaic, Honorific, Colloquial) and sentence-final morpheme choices.
- **Wordplay (88 segments):** Evaluates the adaptation of linguistic humor, onomatopoeia, and visual sound effects created via deconstructed Hangul characters.
- **Hate Speech (56 segments):** Captures emotional intensity and lexical variety in profanity and derogatory expressions.
- **Naming Transcreation (45 segments):** Evaluates world-building name adaptation for character/race names and items.
- **Interjection (39 segments):** Assesses natural realization of exclamations and vocalized reactions.
- **Cultural Reference (33 segments):** Evaluates Korean transcreation of culture-specific Western terms.
- **Gender Stereotyping (20 segments):** Diagnoses gender bias in translation and evaluates neutral expression choices.
- **Idiom & Proverb (16 segments):** Evaluates the concise adaptation of English expressions into culturally equivalent Korean idioms and proverbs.

4 Automatic Evaluation Metrics

We establish a panel of four automated metrics combining reference-based metrics (BLEU, COMET) and reference-free Quality Estimation (QE) metrics (CometKiwi, MetricX). To mitigate the under-estimation tendency of reference-based metrics in game localization—where cultural transcreation is frequent—QE metrics provide complementary signals.

- **BLEU** (Papineni et al., 2002): Precision-based n-gram lexical overlap (↑).
- **COMET-22** (Rei et al., 2022b): Neural reference-based metric fine-tuned on human quality judgments (↑).
- **CometKiwi** (Rei et al., 2022a): Reference-free QE metric predicting quality directly from source and hypothesis (↑).
- **MetricX** (Juraska et al., 2024): mT5-based quality estimation metric predicting error severity (↓).

Because these metrics differ in scale, direction, and reference dependency, raw scores cannot be directly averaged. In §6.1, scores are converted to category-specific ranks for aggregation, while discrepancies between reference-based and QE metrics are analyzed in §6.3.

5 Category-Specific Evaluation Protocols

In this section, we describe a hybrid evaluation pipeline tailored to the characteristics of the four primary categories, employing Claude Sonnet 5 as the judge to mitigate self-preference bias in LLM-as-a-Judge evaluations (see Appendix A for model selection rationale, and Appendix B for a human meta-evaluation validating the judge against a three-annotator panel).

5.1 Terminology & Accuracy (TA)

This category evaluates whether target vocabulary is correctly reflected based on clear ground-truth criteria (canonical terms or polysemous contexts).

Rule-based evaluation: Evaluates translations in two stages. First, after minimal normalization (removing whitespace and punctuation), it checks whether the target vocabulary appears as a substring. If it fails, it proceeds to the second stage, extracting nominal lemmas (common nouns, proper nouns, foreign words) using a morphological analyzer to verify if the target lemma combination is present. Passing either stage grants a score of 1.

LLM-as-a-Judge: Claude Sonnet 5 judges correctness (0 or 1) per segment, tolerating flexible Korean spacing and morphological variations while strictly penalizing semantic drift.

5.2 Persona & Pragmatics (PP)

This category contrasts performance between the baseline and Regime R augmented with contextual metadata to measure the "prompt gap" indicating socio-linguistic blind spots. Following the GEMBA framework (Kocmi and Federmann, 2023), a zero-shot LLM-as-a-Judge protocol using Claude Sonnet 5 is adopted (Zheng et al., 2023). Source meaning and speech acts serve as primary criteria, while reference translations act only as observable examples.

Rubric-based LLM-as-a-Judge: Evaluates meaning preservation, register appropriateness, and naturalness using 5-point mapped labels.

Excessive Over-Persona Handling: Strong expressions supported by context (e.g., urgent military tones) are accepted. Conversely, unsupported emotional additions or extreme registers (e.g., overly archaic endings) are penalized as unwarranted meaning additions. Evaluators prioritize overall discourse naturalness over single sentence-final endings.

5.3 Technical Robustness (TR)

This category focuses on pass/fail compliance with game engine and UI constraints, applying a conservative LLM-as-a-Judge exclusively to segments that pass prior rule-based verification to binarily verify meaning preservation while preventing formally compliant semantic distortion.

Step 1: Rule-based Verification

- **Placeholder Integrity:** It checks for missing or corrupted tokens such as `%s` or `$var`, while permitting particle attachments (e.g., `%s이(가)`) following Korean localization conventions and deferring token reordering verification to the second-stage judgment.
- **Tag Integrity:** For paired closing tags such as `<i>`, `<b>`, and `<span>`, it inspects inventory matching (type and count) and stack-based well-formedness, while prohibiting empty spans (`<i></i>`).
- **Length Constraint:** Computes terminal cell width based on East Asian character widths. Scores are defined as:

$$\text{width_score} = \min\left(1,\ \frac{\theta \cdot w_{\text{base}}}{w_{\text{hyp}}}\right),\quad \theta = 1.75$$

where the tolerance multiplier θ is set based on the minimum ratio passing all reference texts (50/50).

Step 2: Binding Check For segments that pass rule-based verification, an LLM-as-a-Judge evaluates whether the elements properly correspond to the source argument structure without considering

surface fluency or reordering, assigning a binary score (`correct_binding` = 1 or 0).

5.4 Cultural Transcreation (CT)

CT measures localization adaptation tailored to Korean gamers, moving beyond literal translation. All sub-categories are evaluated using Claude Sonnet 5 via a 5-point LLM-as-a-Judge rubric.

- **Wordplay, Interjection:** Evaluates acoustic/morphological reproduction and creativity, penalizing copy bias, transliteration bias, and expression flattening.
- **Naming Transcreation, Cultural Reference, Idiom & Proverb:** Evaluates idiomaticity and genre appropriateness to evaluate context-aware localized adaptation.
- **Speech Style Adaptation:** Evaluates macro-styles (Step 1, weighted 40%) and marker realization (Step 2, weighted 60%) in a hierarchical dual structure.
- **Hate Speech & Gender Stereotyping:** Evaluates source-grounded equivalence rather than reference match, since reference translations often contain sanitization or gender biases.

6 Experiment Results and Analysis

This section presents an in-depth evaluation of the 27 translation systems using KUEST. We report automated metric results in §6.1, detail category-specific evaluations in §6.2, and discuss key insights in §6.3.

6.1 Metrics

Table 3 presents the performance of 27 evaluated systems across KUEST's four primary categories using automatic metrics (BLEU, COMET, CometKiwi, and MetricX), along with their overall ranks calculated across all 16 category-metric combinations.

Overall Rank: SRPOL achieved the top overall position among all 27 systems with an average rank of 6.44. It was closely followed by GPT-5.5 (6.94), GoogleTranslate (7.12), and Wayfinder (7.50), together forming the top tier. The lowest-ranked Gemini 3.1 Pro suffered from reasoning artifacts leaking into the output, resulting in translation lengths roughly 16 times that of the reference and severely degrading its evaluation scores.

Divergence Between Standard and QE Metrics:

- **Lexical & Neural Metrics(BLEU/COMET):** Systems such as Dubformer (TA BLEU: 37.9, COMET: 87.0), GPT-5.5 (PP BLEU: 17.7, COMET: 87.5; TR BLEU: 29.3, COMET: 82.4), and TRIVE (CT BLEU: 17.9) recorded the highest scores, demonstrating strong lexical and semantic similarity to reference translations.
- **Reference-Free QE Metrics (CometKiwi/MetricX):** SRPOL consistently outperformed other systems across all four categories in CometKiwi (TA: 82.9, PP: 85.8, TR: 78.5, CT: 81.7) and three categories in MetricX (TA: 2.37, PP: 2.53, CT: 2.77). This demonstrates SRPOL's distinct strength in standalone translation completion and error control, independent of lexical similarity to reference translations.

Category-Specific Characteristics: In the TR (Technical Robustness) category, GoogleTranslate recorded the lowest error rate with a MetricX score of 3.27. In the CT (Cultural Transcreation) category, rank fluctuations between traditional metrics (BLEU/COMET) and QE metrics (CometKiwi/MetricX) were most pronounced, confirming a substantial divergence in metric perspectives when evaluating game-specific transcreation elements.

6.2 Evaluation Results and Analysis by Category

6.2.1 Terminology & Accuracy (TA)

In this category, we deeply analyze and compare the performance of 27 translation systems from two perspectives: string-matching-based rule validation and flexible context-understanding-based LLM-as-a-Judge validation.

Limitations of the Rule-Based Approach: Rule-based validation strictly recognizes exact string matches with the designated glossary. While Gemini 3.1 Pro achieved the highest score on the Rule axis by most conservatively and accurately reflecting the provided glossary without modification, it exhibits a limitation of failing Rule evaluation when flexible contextual paraphrasing or inflectional variations are applied.

Synonym and Paraphrase Acceptability in LLM-as-a-Judge: This approach evaluates translations based on semantic validity rather than strict keyword inclusion, assessing whether

polysemous expressions carry the correct contextual sense. Systems such as Dubformer and Qingqiu-MT-9B actively utilized natural synonyms and paraphrases tailored to the gaming context despite slightly lower surface-level lexical match rates, thereby earning higher scores in the LLM-as-a-Judge evaluation, which recognizes contextual validity.

6.2.2 Persona & Pragmatics (PP)

We evaluated 388 segments (totaling 5,238 corresponding pairs across 27 systems) under Baseline and Regime R environments using a 5-point rubric. While 51.6% (2,705 cases) received identical scores across both conditions, among the remaining 2,533 divergent cases, Baseline outperformed Regime R in 1,309 cases compared to 1,224. Although 19 out of 27 models exhibited a negative prompt gap, the overall average scores were identical across both conditions (3.786 each). This is a statistical artifact caused by Gemini 3.1 Pro, whose reasoning leakage (+2.660) masked the negative gaps of other models. Excluding this outlier, the Baseline average (3.871) exceeded that of Regime R (3.769), with the Baseline achieving a win rate of 55.4%.

Over-integration of Speaker Information: The reduction of neutral translations (scored as 3) from 381 to 248 cases indicates that models actively incorporated the provided metadata; however, this shift migrated toward 2-point scores rather than 5-point scores (increasing from 629 to 878 cases). The primary reasons for penalty deductions were tone/formality mismatch (34.6%), semantic and speech-act errors (32.5%), and excessive archaic or over-referential honorific registers (12.8%). This suggests that models treated social status tags as deterministic lookup tables for sentence endings, causing the original semantic meaning and speech intention to be subordinated to the tags. For instance, a commoner's utterance *"What be your pursuits?"* shifted from "당신이 추구하는 바는 무엇인가요?" (*"What is it that you pursue?"*, scored as 4) to "네 일은 뭐야?" (*"What's your job?"*, scored as 3). Similarly, for Lady Jessene, a royal female speaker, *"What? Why would you doubt me?"* shifted from "뭐? 왜 날 의심하는 거야?" (*"What? Why do you doubt me?"*, scored as 4) to "뭐라? 어찌 나를 의심하는가?" (*"What sayest thou? Wherefore dost thou doubt me?"*, scored as 2), over-applying a gender-neutral archaic written style.

Utterance-Length-Dependent Effects: In 9.7% (507 cases) of the total benchmark, outputs from both conditions were completely identical. For example, *"Harper, no!"* yielded identical translations across 19 out of 27 systems, with short utterances limiting metadata capacity and driving convergence between conditions.

Model Disparity in Metadata Integration: Within the same model family, GPT OSS 20B degraded significantly by -0.418, whereas the 120B variant maintained its performance at $+0.015$. Similarly, among 9B translation-specialized models, Tower 9B ($+0.067$) and Qingqiu-MT-9B (-0.335) demonstrated opposite trajectories. Since all four systems generated substantial output differences across conditions, this disparity cannot be attributed to a simple failure to follow prompt instructions. Therefore, this gap stems not merely from parameter scale or translation specialization, but rather from intrinsic differences in model capability to sophisticatedly select and integrate provided contextual metadata.

6.2.3 Technical Robustness (TR)

We evaluated the ability of models to adhere to technical constraints across 507 segments.

Overwhelming Constraint Compliance of Top-Tier Models: As shown in Figure 4, commercial frontier models and domain-specific models demonstrated strong performance, recording overwhelming constraint compliance rates of over 95%. GPT-5.5 achieved the highest score with 504 passes (99.4%), followed by TRIVE (502, 99.0%), Wayfinder (500, 98.6%), HW-TSC (496, 97.8%), and Qingqiu-MT-9B (496, 97.8%).

Difficulty Gaps Across Evaluation Types: Among the four evaluation types, the largest performance gap between models appeared in tag preservation (average pass rates: placeholder-consistency 89.3%, placeholder-diverse 90.1%, tag preservation 71.1%, length constraint 91.7%). In particular, out of 95 tag preservation items, Tower 9B and Qwen-3.5-9B passed only 15 items, Qwen-3.6-27B passed 41, and GoogleTranslate passed 57, showing a stark contrast between lower- and upper-tier models, whereas top-tier models successfully passed most tag and length constraint items.

Major Failure Types: Tag deletion, placeholder omission/addition, and the generation of extraneous explanations were repeatedly observed. GoogleTranslate and the Qwen series exhibited numerous cases of $< i >$ and $< b >$ tag removal; GPT OSS 20B showed placeholder omissions and the arbitrary introduction of positional placeholders; and Gemini 3.1 Pro generated alternative translations and explanations alongside the output, leading to added placeholders and tags as well as length constraint violations. In contrast, GPT-5.5 successfully passed all formatting rules, recording only 3 semantic binding errors.

6.2.4 Cultural Transcreation (CT)

We evaluated cultural transcreation performance—going beyond literal translation to recreate the cultural background and nuances of the source text to fit the target culture—highlighting performance gaps across subcategories and style flattening in Hate Speech.

Performance Gaps Across Sub-Categories: Performance patterns varied significantly across sub-categories. Top models achieved near-perfect scores in Idiom/Proverb (top: 4.81), Wordplay (top: 4.80), and Interjection (top: 4.74). Conversely, scores remained in lower ranges for Cultural Reference (top: 2.91), Naming Transcreation (top: 3.02), and Speech Style Adaptation (top: 3.28). This suggests that for open-ended tasks governed by strict contextual constraints, fine-grained multi-criteria evaluation inherently leads to more conservative score distributions.

Style Flattening in Adaptive and Sensitive Contexts: Even when the core meaning of the source text was preserved, a widespread tendency to soften the intensity of profanity, derogatory expressions, and character traits was frequently observed. While this style flattening is also observable in Speech Style Adaptation, it becomes particularly pronounced in the Hate Speech category. Out of 1,512 total outputs in Hate Speech, 408 (27.0%) were classified as `bleached`, with certain items showing the same softening trend across up to 26 systems, indicating a widespread model-wide phenomenon (See Table 12.) Meanwhile, given that the evaluator, Claude, is based on Constitutional AI (Bai et al., 2022), potential intervention from safety alignment cannot be ruled out.

6.3 Discussion

The evaluation results of this study suggest that automatic metrics have clear limitations in diagnosing compliance with key evaluation aspects. By contrasting the results of §6.1 and §6.2, we derive three key academic implications as follows.

Distortion of Rankings in Automated Evaluation Metrics: Automated evaluation metrics pose a significant risk of distorting actual performance. For instance, SRPOL, which ranked 1st overall in §6.1 due to the illusion of reference-free QE metrics, fell to 15th (3.29/5) in the in-depth CT evaluation (§6.2.4). Conversely, UvA-MT and TRIVE, which ranked 17th and 10th respectively under automated metrics, bounced back to 2nd (3.92/5) and 1st (3.96/5) in CT evaluation. This demonstrates that automated evaluation metrics—including surface-level similarity measures like BLEU—fail to fully capture creative translation quality. Furthermore, it reveals that KUEST effectively exposes domain constraint adherence that these automated metrics overlook.

Residual Category-Specific Variance: Strong systems tend to be strong everywhere, and rank inversions across the four KUEST categories are the exception rather than the rule. Against that backdrop, we reveal residual capability fragmentation—category-specific variance that survives the shared general-competence factor and inverts top-tier rankings. Dubformer ranks 1st in TA (75.2%) but 8th in TR, and TRIVE 1st in CT (3.96/5) but 5th in TA. KUEST categories are positively correlated ($\rho = 0.600$–0.953; see Table 13), and we do not claim independence: the sharing is uneven, with the weakest association observed between TA and TR ($\rho = 0.600$), and the rank inversions above occurring along these more weakly correlated category pairs. Our claim concerns residual variance sufficient to invert top-tier standings, where the cost is incurred—SRPOL ranks 1st on automatic metrics yet 15th in CT. Nor is the effect confined to the weakest-sharing pairs: HW-TSC falls from 4th in TR to 16th in CT even though that pair is closely aligned ($\rho = 0.861$). Taken together, these results show why a single composite score can hide which of several near-equal systems is right for a given localization task; dissecting performance by category, as KUEST does, is what makes that difference visible.

System Landscape and Tier Analysis of 27 Systems: The comprehensive positioning analysis in Figure 1 clearly reveals distinct characteristics across groups:

- **GenMT Participants:** The majority of translation-specialized models (e.g., TRIVE, Wayfinder, and Lumen) heavily populated the upper-right quadrant (Both strong), exhibiting consistently robust performance across evaluation axes.
- **Commercial & Proprietary Models:** Except for GPT-5.5, the remaining commercial offerings are predominantly MT-specialized Models (e.g., Dubformer and GoogleTranslate), which maintained high positions across both axes regardless of model tier. Conversely, Gemini 3.1 Pro recorded the lowest evaluation due to the leakage of reasoning tokens. This suggests that robust overall translation capability and decoding stability are far more decisive than model scale alone in balancing the dual demands of game translation and adaptation: strict rule compliance and flexible contextual transcreation.
- **Open-Weights LLM Baselines:** With a few high-capacity exceptions such as DeepSeek V4 Pro and Gemma 4-31B, the vast majority of general open-weights baselines remained in the lower-left quadrant (*Both weak*). This underscores the inherent limitations of general-purpose LLMs without explicit translation-specific tuning or game domain adaptation.

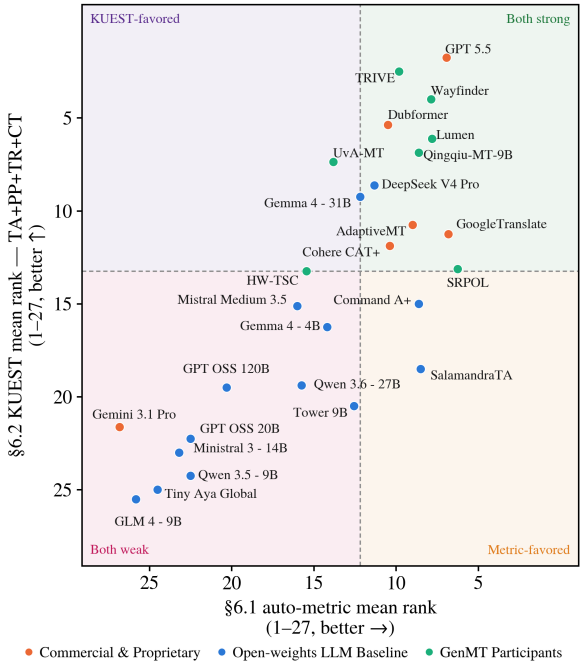


Figure 1: Tier map of the 27 systems. The x-axis represents the average rank based on automatic metrics (§6.1), and the y-axis represents the average rank based on **KUEST** (§6.2) (ranging from 1 to 27, with the top-right indicating better performance). Dotted lines divide the four quadrants based on the median rank of each axis.

7 Conclusions

This study establishes three key contributions through KUEST, a benchmark designed to diagnose the complex constraints of game localization:

1. **Demonstration of the Fallacy of Metrics:** We show that standard machine translation and general QE metrics fail to represent performance in cultural adaptation (CT) and technical rule compliance (TR), inducing severe ranking distortions.

2. **Residual Category-Specific Variance:** We identified that beyond a shared general-competence factor, category-specific variance remains large enough to invert top-tier rankings, as model strengths and limitations manifest depending on the evaluation dimension.

3. **Identification of the Prompt Gap:** By observing performance degradation in numerous systems when incorporating speaker and relationship metadata, we revealed persistent sociolinguistic blind spots in LLMs' conditional instruction-following capabilities.

8 Limitations and Future Work

Despite these meaningful contributions, this study has limitations in terms of genre and text type diversity, as it focused exclusively on a single English-Korean language pair and two open-source games (CDDA and Wesnoth). In future work, we aim to incorporate license-compatible open-source games to expand **KUEST** into a genre-comprehensive general game test suite. Furthermore, we intend to extend the target language pairs to multilingual settings to rigorously verify whether the dataset construction methodology and evaluation framework proposed in this work are equally applicable across diverse linguistic environments.

Acknowledgements

The authors would like to thank **Yongrae Jo**[4] for his valuable advice and insightful feedback throughout this research. We also thank **Pseudo Lab**[5], a non-profit AI/ML research community, for providing the collaborative environment that brought the authors together and enabled this research. This research was supported by the **"Open Source AI · SW Developer and Community Support Program"** funded by the **Ministry of Science and ICT (MSIT), Republic of Korea.**

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Drain Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 28 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Miguel Ángel Bernal-Merino. 2015. *Translation and Localisation in Video Games: Making Entertainment Software Global*. Routledge, New York.

Marta R. Costa-jussà, Emily Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Greg Hanneman and Georgiana Dinu. 2020. How should markup tags be translated? In *Proceedings of the Fifth Conference on Machine Translation* (*WMT*), pages 1160–1173, Online. Association for Computational Linguistics.

Damien Hansen and Pierre-Yves Houlmont. 2022. A snapshot into the possibility of video game machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas* (*Volume 2: Users and Providers Track*), pages 257–269, Orlando, USA. Association for Machine Translation in the Americas.

[4] dreamgonfly@gmail.com

[5] https://pseudo-lab.com/

Yejin Hwang, Young-Bum Kim, and Kyomin Jung. 2021. Context-aware neural machine translation for korean honorific expressions. *Electronics*, 10(13):1589.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Juraj Juraska, Lev Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation* (*WMT*), pages 451–457, Miami, Florida, USA. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 13th International Conference on Computational Linguistics* (*COLING*), Helsinki, Finland.

Tom Kocmi, Eleftherios Eleftheriadis, Christian Federmann, Barry Haddow, Matthias Huck, Marcin Junczys-Dowmunt, Rebecca Knowles, Samuel Larkin, Christof Monz, Makoto Morishita, and Keisuke Sakaguchi. 2024. Findings of the 2024 conference on machine translation (WMT24): The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation* (*WMT*), pages 1–54, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA: GPT-based metric for translation quality assessment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 15160–15174, Singapore. Association for Computational Linguistics.

Tom Kocmi, Markus Freitag, Ekaterina Artemova, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, and 9 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation* (*WMT*), pages 355–413, Suzhou, China. Association for Computational Linguistics.

Surafel M. Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation* (*IWSLT*).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German–English machine translation

based on a test suite. In *Proceedings of the Third Conference on Machine Translation* (*WMT*): *Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Carme Mangiron. 2016. Games without borders: The cultural dimension of game localisation. *Hermēneus*, 18:187–208.

Carme Mangiron and Minako O'Hagan. 2006. Game localisation: Unleashing imagination with 'restricted' translation. *The Journal of Specialised Translation*, (6):10–21.

Naman Moghe, Arefa Mutaser, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Maria Nădejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems* (*NeurIPS*), volume 37.

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

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long Papers*), pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation* (*WMT*)*, Volume 2: Shared Task Papers*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, José G. C. De Souza, Duarte M. Alves, Chrysoula Zerva, Marcos Treviso, Ana Guerberof Arenas, and André F. T. Martins. 2022a. Cometkiwi: Unsupervised Metrics for Machine Translation Evaluation. In *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Ana Guerberof Arenas, Duarte M. Alves, Chrysoula Zerva, José G. C. De Souza, and André F. T. Martins. 2022b. COMET-22: Uncertainty-aware Metric for Machine Translation Evaluation. In *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Dimitar Shterionov, Eva Vanmassenhove, and Andy Way. 2020. NEUROMANTICS: Neural machine translation drops linguistic richness. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (*CoNLL*), pages 376–388, Online. Association for Computational Linguistics.

Ho-Min Sohn. 1999. *The Korean Language*. Cambridge University Press, Cambridge.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Lisa Yankovskaya, and Anna Kozlova. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation* (*WMT*), pages 684–702, Online. Association for Computational Linguistics.

Harry Zhang, Qiang Wang, Bing Qin, Zemin Shi, Huaiyu Wang, and Mu Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 6029–6042, Toronto, Canada. Association for Computational Linguistics.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems* (*NeurIPS*), volume 36, pages 46595–46623.

A Evaluator Model Selection Rationale

The theoretical rationale for selecting Claude Sonnet 5 as the primary evaluator model in our LLM-as-a-Judge protocol is as follows:

Exclusion of Self-Preference Bias While GPT-4-class models have served as the de facto standard evaluators in machine translation literature (Kocmi and Federmann, 2023), the systems benchmarked in this leaderboard are constructed from 27 translation systems that officially participated in the WMT Shared Task (Kocmi et al., 2024). Prior studies demonstrate that LLM evaluators exhibit a marked "self-preference" bias, systematically favoring outputs generated by models of the same architecture or model family (Panickssery et al., 2024; Zheng et al., 2023). Notably, the official WMT submission roster contains no models from Anthropic's Claude family. By appointing Claude —a frontier model completely independent from the WMT candidate pool—as the final judge, we reduce the risk of potential evaluation contamination and self-preference biases that could arise when employing GPT, Gemini, or DeepSeek models. Furthermore, this choice aligns with recent trends in the WMT metric pipeline that advocate diversifying evaluation frameworks beyond reliance on a single GPT judge (Kocmi et al., 2025).

Rationale for Claude Sonnet 5 Selection Within the Claude family, Sonnet was selected over the highest-tier Opus model to balance cost-efficiency and throughput across leaderboard-scale evaluations. Claude Sonnet 5 operates as a highly capable "workhorse" model equipped with strong reasoning capabilities, consistent with the standard practice of employing this model tier for large-scale benchmark evaluations. Additionally, a qualitative human review conducted on a sample subset revealed that the judgment quality between Opus and Sonnet remained substantially comparable, serving as supporting evidence for prioritizing Sonnet.

B Human Evaluation for LLM-as-a-Judge Validation

B.1 System Selection

To validate the reliability of the LLM-as-a-Judge (Claude Sonnet 5), we conduct a human meta-evaluation comparing judge scores with human ratings. Evaluating all 27 systems would require substantial annotation cost. We therefore select three representative systems based on three criteria: coverage of the quality spectrum, validity of translation outputs, and diversity of MT system families.

Coverage of the quality spectrum. Rank-based agreement can be attenuated under range restriction when evaluated systems occupy a narrow quality range. We therefore select one system from each of the top, middle, and bottom tiers of the overall KUEST ranking (Table 15): GPT-5.5 (1st), GoogleTranslate (12th), and Qwen-3.5-9B (25th). This selection spans a wide range of system quality, providing sufficient variation to assess agreement between LLM-as-a-Judge scores and human judgments.

Valid translation outputs. We exclude systems whose low scores are driven primarily by output collapse rather than translation quality. Gemini-3.1-Pro leaks reasoning traces, inflating its outputs to roughly 16 times the reference length (§6.1), while GLM-4-9B leaks reasoning in about 5% of its segments. Including such systems would allow humans and the judge to agree simply because the outputs are unusable, without testing whether the judge can discriminate along the rubric. All three selected systems produce valid translations; even the lowest-ranked, Qwen-3.5-9B, shows no reasoning leakage and only about 1.5% polite refusals on profanity items. These cases therefore provide meaningful evidence of rubric-level judging rather than output-collapse detection.

Coverage of distinct MT system families. The three systems represent distinct system families: a frontier LLM (GPT-5.5), a commercial machine translation system (GoogleTranslate), and an open-weight LLM (Qwen-3.5-9B). This diversity reduces the possibility that the meta-evaluation is specific to a single output style or model family.

B.2 Sampling Procedure

We conducted stratified sampling across the 15 evaluation subcategories. Each subcategory received a minimum allocation of three instances to ensure coverage of low-resource phenomena, while the remaining samples were allocated proportionally to the original stratum sizes. This resulted in a 100-instance sample balancing phenomenon coverage and preservation of the overall benchmark distribution.

B.3 Human Evaluation Setup

To validate the reliability of our LLM-as-a-Judge (LaaJ), we conducted a human meta-evaluation using the same evaluation rubric and scoring criteria as the LaaJ. Specifically, three human evaluators independently assessed the translations produced by the three selected systems on a stratified sample of 100 segments from the 1,796 segments in KUEST. The shared rubric and scoring criteria enable a direct comparison between human judgments and automated judge scores.

The human evaluation covered the same evaluation categories and subcategories as the LaaJ, with one exception. The *Length Constraint* subcategory under Technical Robustness (TR) was excluded from human evaluation because this subcategory was not evaluated by the LaaJ. Thus, the human meta-evaluation was restricted to dimensions for which corresponding LaaJ scores were available.

Each of the three evaluators independently assigned rubric-based scores to the sampled translations. These individual ratings were used both to measure inter-annotator agreement among human evaluators and to assess the agreement between human judgments and the LaaJ scores.

B.4 Agreement Measures

Excluding *Length Constraint* removes 15 of the 300 segment–system pairs, leaving 285 rated items —108 on the binary Pass/Fail axes (TA, TR) and 177 on the 1–5 rubric axes (PP, CT). We report Krippendorff's α—nominal on the binary axes and ordinal on the rubric axes—together with mean pairwise Spearman correlations and the two-way $\mathrm{ICC}(2, \cdot)$ for the fixed panel of three annotators.

B.5 Inter-Annotator Agreement

Human annotators agree closely on both scales (Table 2): $\alpha = 0.828$ on the binary axes and $\alpha = 0.807$ on the rubric axes, with $\mathrm{ICC}(2, 1) = 0.820$ for a single annotator and $\mathrm{ICC}(2, 3) = 0.932$ for the three-annotator mean. The human ratings therefore provide a stable reference for comparison with the LaaJ.

B.6 Agreement with the LLM-as-a-Judge(LaaJ)

The LaaJ matches the human majority verdict on 82.4% of binary items and correlates with the three-annotator mean at $\rho = 0.773$ (MAE $= 0.61$; 83.6% of scores within one rubric point). Its mean pairwise correlation with individual human annotators is $\rho = 0.736$, somewhat below the human–human correlation of $\rho = 0.812$. The LaaJ score falls within the range of the three human annotators on 74.0% of items, compared with 76.3–81.9% pairwise range coverage among the human annotators. At the system level, the LaaJ preserves the same ranking as the human means, with slightly higher absolute scores (human means: 4.02/3.46/2.42; LaaJ: 4.14/3.70/2.49).

Axis	n	Human α	Human	LaaJ
Binary axes (agreement with majority verdict)				
TA	48	0.959	97.9–100%	93.8%
TR	60	0.622	83.3–100%	73.3%
All	108	0.828	89.8–100%	82.4%
Rubric axes (mean pairwise Spearman ρ)				
PP	54	0.758	0.768	0.742
CT	123	0.836	0.840	0.721
All	177	0.807	0.812	0.736

Table 2: Human and LaaJ agreement. α is Krippendorff's α over the three annotators (nominal above, ordinal below). The last two columns report agreement with the human majority verdict in the upper panel, and mean pairwise Spearman correlations in the lower.

The only weak axis is TR, driven by placeholder reordering segments ($\alpha = 0.201$). The low coefficient contrasts with high human agreement: the three annotators gave unanimous ratings on 29 of 42 items, and all 13 disagreements were single-annotator dissents, 10 from the same annotator. This suggests a difference in annotator strictness rather than substantial item-level ambiguity. Accordingly, the LaaJ's TR results should be interpreted against this low-reliability human reference.

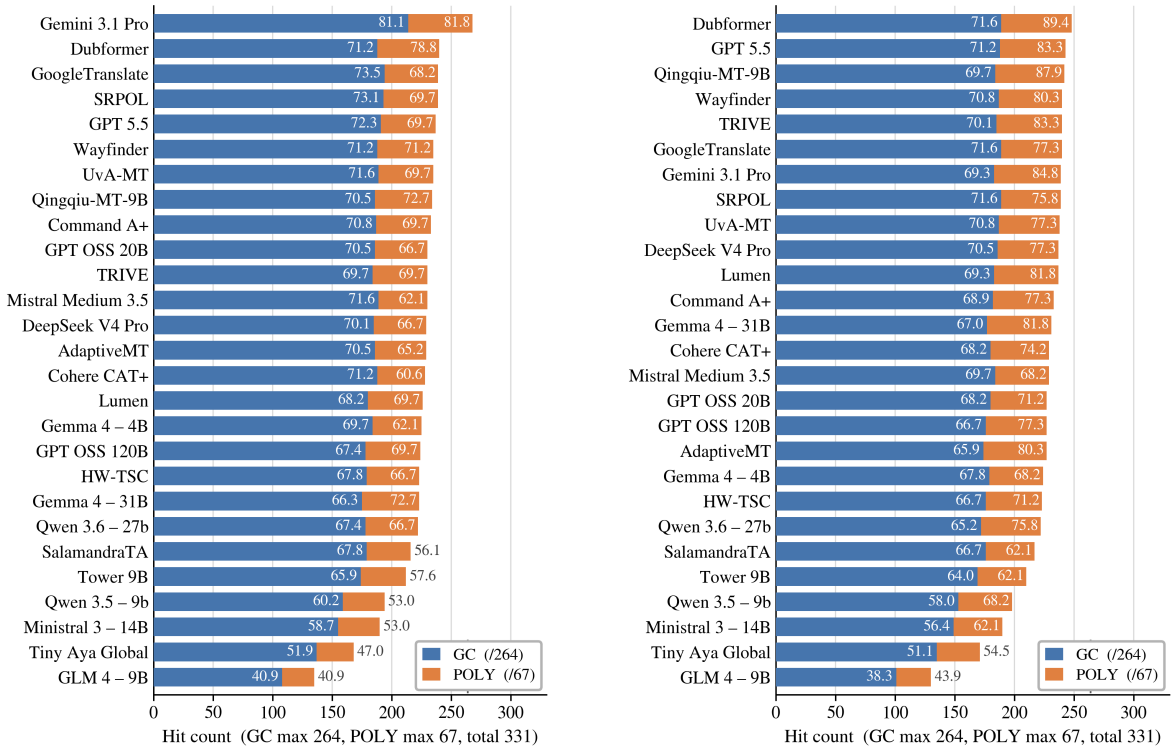
B.7 Conclusion

Because no gold standard exists for rubric-based translation quality, human agreement provides the appropriate empirical reference for evaluating the LaaJ. Measured against it, the LaaJ agrees substantially with human judgments while showing somewhat greater variability than an individual annotator ($\rho = 0.736$ vs. 0.812; inside the annotators' range on 74.0% of items vs. 76.3–81.9%), and it reproduces the human system ordering exactly. We therefore consider LaaJ scores reliable for the system-level and category-level comparisons reported in this paper, while not treating them as ground truth for individual segments, where a residual error of roughly one rubric point remains (MAE $= 0.61$).

C Category-wise Evaluation Results and Visualizations

System	TA				PP				TR				CT				Overall
	BLEU	COMET	CKW	MetX	BLEU	COMET	CKW	MetX	BLEU	COMET	CKW	MetX	BLEU	COMET	CKW	MetX	
SRPOL	31.2	86.0	**82.9**	**2.37**	13.3	87.3	**85.8**	**2.53**	25.4	80.8	**78.5**	3.36	13.0	76.7	**81.7**	**2.77**	**6.44**
GPT 5.5	35.5	86.6	81.4	2.73	**17.7**	**87.5**	84.9	3.12	**29.3**	**82.4**	76.9	3.67	17.3	**78.7**	79.1	3.17	6.94
GoogleTranslate	34.4	85.1	82.4	2.41	15.5	87.4	85.2	2.72	24.3	79.3	76.9	**3.27**	14.4	76.5	80.3	2.81	7.12
Wayfinder	35.8	86.5	81.6	2.75	15.5	87.0	84.7	2.94	28.4	81.9	77.0	3.60	15.9	78.2	79.4	3.06	7.50
Command A+	31.9	85.5	82.5	2.38	13.4	85.9	85.5	2.75	22.2	79.7	77.9	3.32	11.8	76.2	81.4	2.78	8.62
Qingqiu-MT-9B	33.3	86.2	81.8	2.71	15.6	86.6	84.9	2.98	28.2	**82.4**	76.6	3.72	14.6	77.3	79.6	3.15	8.69
SalamandraTA	26.2	83.9	82.7	2.53	14.5	87.3	85.8	2.90	25.5	80.9	76.8	3.57	12.6	76.9	81.3	2.94	8.81
Lumen	34.6	85.9	81.8	2.56	15.9	86.6	84.6	3.12	27.3	81.3	76.7	3.70	16.2	77.6	79.3	3.12	9.00
AdaptiveMT	31.5	85.2	82.0	2.51	15.8	86.9	84.9	3.03	26.3	80.4	76.3	3.68	15.5	77.3	79.7	3.04	9.12
TRIVE	36.0	86.0	81.3	2.76	16.2	87.3	84.9	3.00	27.8	82.2	75.7	3.84	**17.9**	78.5	78.4	3.21	9.38
Cohere CAT+	28.6	84.2	82.0	2.61	15.1	86.1	84.6	2.91	27.3	81.2	76.7	3.47	14.3	77.1	79.7	2.95	10.12
Dubformer	**37.9**	**87.0**	81.1	2.78	16.2	86.5	84.1	3.16	28.5	81.7	76.4	3.81	17.0	77.5	79.0	3.21	10.62
DeepSeek V4 Pro	34.9	85.1	81.3	2.71	17.1	86.6	84.9	3.17	26.4	80.4	76.3	3.81	17.2	77.5	79.1	3.30	11.00
Tower 9B	29.8	83.9	81.8	2.60	11.7	85.4	85.4	2.91	20.4	78.2	77.2	3.53	10.7	75.1	80.8	3.00	12.69
Gemma 4 - 31B	35.2	85.7	81.3	2.64	17.3	86.9	84.6	3.16	26.9	80.2	75.1	3.87	16.3	76.3	78.4	3.41	12.88
Gemma 4 - 4B	30.2	84.5	80.9	2.51	15.2	85.5	85.0	3.01	22.4	77.8	75.6	3.56	15.0	76.0	79.3	3.18	13.06
UvA-MT	30.7	85.1	80.7	2.54	14.3	85.4	83.9	3.11	24.7	81.1	76.5	3.64	15.0	77.7	78.4	3.25	13.62
Qwen 3.6 - 27b	31.8	84.7	81.7	2.59	11.7	85.4	84.7	3.33	22.8	79.0	76.1	3.71	12.0	75.6	79.8	3.36	15.50
HW-TSC	28.0	81.8	80.6	2.72	13.9	86.3	84.7	3.09	26.2	79.6	75.3	3.58	13.5	74.0	78.9	3.20	15.62
Mistral Medium 3.5	31.1	84.0	81.0	2.73	14.6	85.5	84.5	3.28	22.2	78.6	76.1	3.59	13.6	75.4	79.1	3.20	15.94
GPT OSS 120B	24.9	81.5	80.9	3.19	11.6	84.0	84.6	3.16	20.2	77.1	75.8	3.83	10.4	73.2	78.8	3.35	20.38
GPT OSS 20B	27.3	82.3	79.8	2.93	10.2	82.8	83.5	3.27	19.5	76.6	75.4	3.89	10.1	73.2	78.2	3.44	22.38
Qwen 3.5 - 9b	26.0	81.4	79.6	2.89	10.3	82.9	83.5	3.46	22.9	77.0	74.8	3.96	10.3	73.3	76.3	3.64	22.50
Ministral 3 - 14B	22.0	78.8	79.2	2.96	12.3	83.1	83.2	3.62	15.3	72.9	73.0	4.16	12.0	73.7	78.8	3.47	23.00
Tiny Aya Global	23.6	77.9	76.8	3.40	12.2	81.4	81.0	3.67	17.2	74.6	72.7	4.03	9.8	70.1	76.8	3.80	24.38
GLM 4 - 9B	19.8	75.7	74.9	4.09	10.6	79.4	79.2	4.39	15.9	73.8	71.9	4.54	8.3	68.8	74.0	4.83	25.81
Gemini 3.1 Pro	11.6	60.2	61.2	5.25	2.5	53.7	58.3	6.24	16.5	65.6	63.7	5.39	3.7	53.2	57.2	6.13	26.88

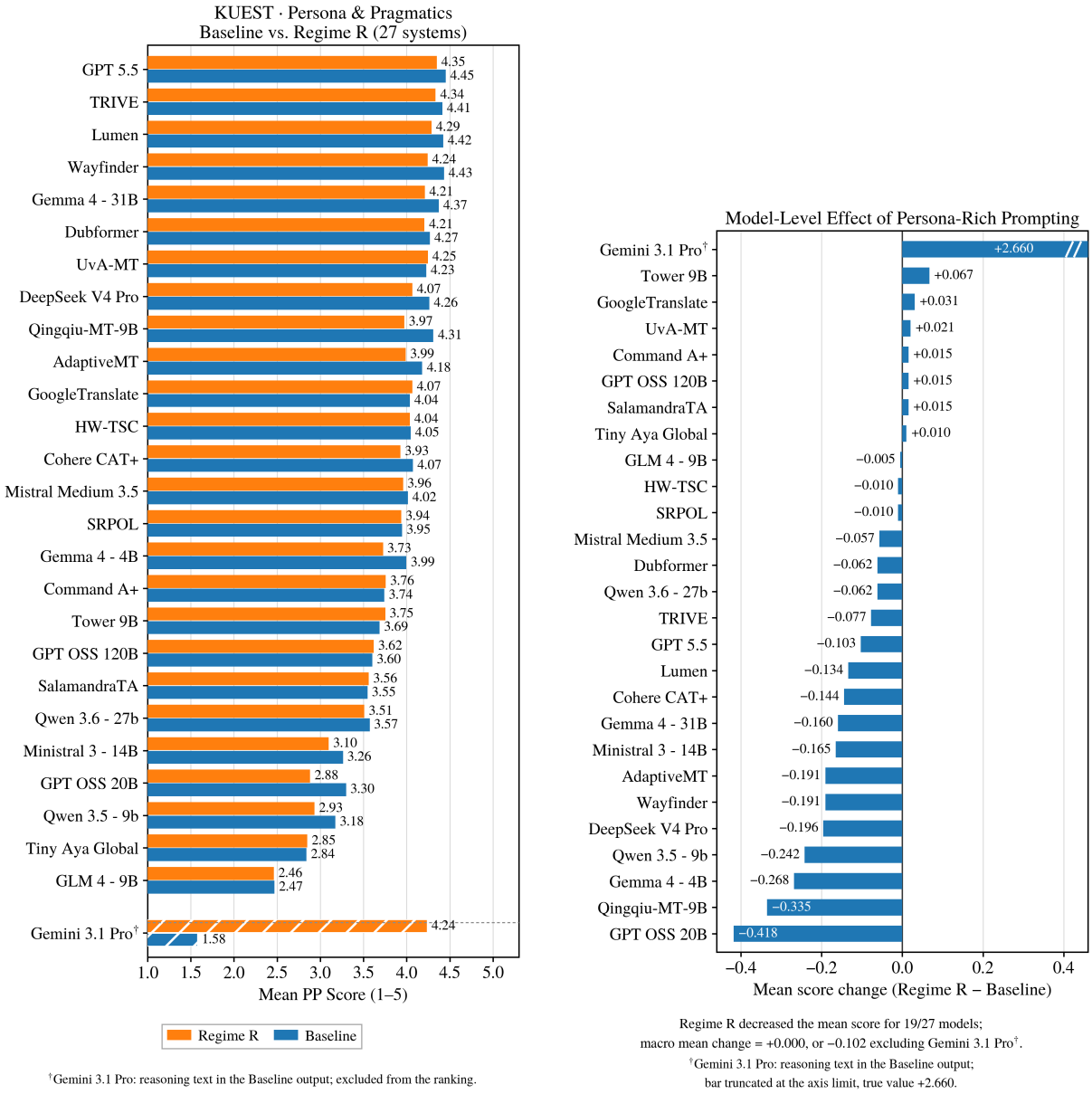
Table 3: BLEU/COMET/CometKiwi(CKW)/MetricX(MetX) by category (TA/PP/TR/CT), 27 systems. MetricX is an error metric (lower is better); all others are higher-is-better. Bold = best in column. Overall = mean rank across the 16 (metric, category) columns (1–27, lower is better; ties receive the average rank).



(a) Rule-based Leaderboard

(b) LLM-as-a-Judge Leaderboard

Figure 2: Terminology & Accuracy (TA) evaluation results across 27 systems. Stacked bar charts show hit counts and percentages for Glossary Compliance and Polysemy, sorted by total hits.



(a) Baseline vs. Regime R Average Scores

(b) Prompt Effectiveness (Regime R − Baseline)

Figure 3: Evaluation results and prompt effectiveness in the Persona & Pragmatics (PP) category across 27 valid systems. (a) Average scores comparison between Baseline and Regime R. (b) Prompt effectiveness by system, where positive values (+) indicate the superiority of Regime R and negative values (-) indicate the superiority of the Baseline. A negative prompt gap—representing the dominance of the basic prompt—was observed across 19 models in total.

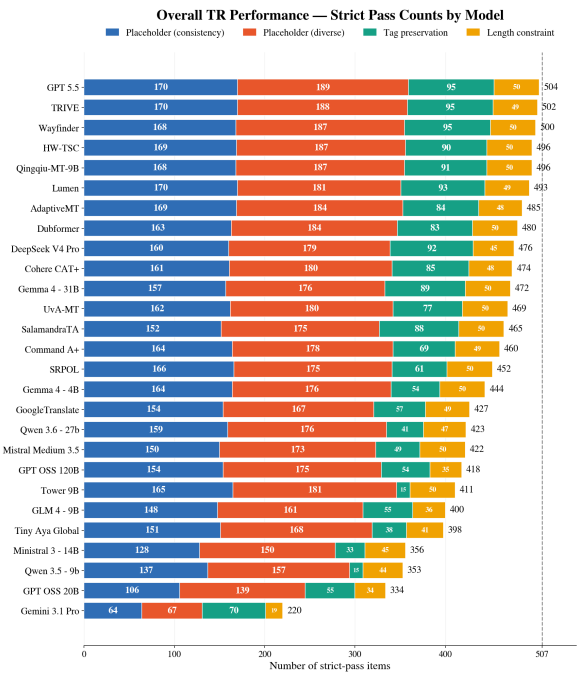


Figure 4: Cumulative bar chart of the number of items passing the strict conditions per model in the Technical Robustness (TR) category (Maximum score: 507). Each segment represents the sum of the number of strictly passed items for the two-stage checks—Placeholder (consistency), Placeholder (diverse), and Tag preservation—along with the number of items complying with the length constraint (width_score = 1).

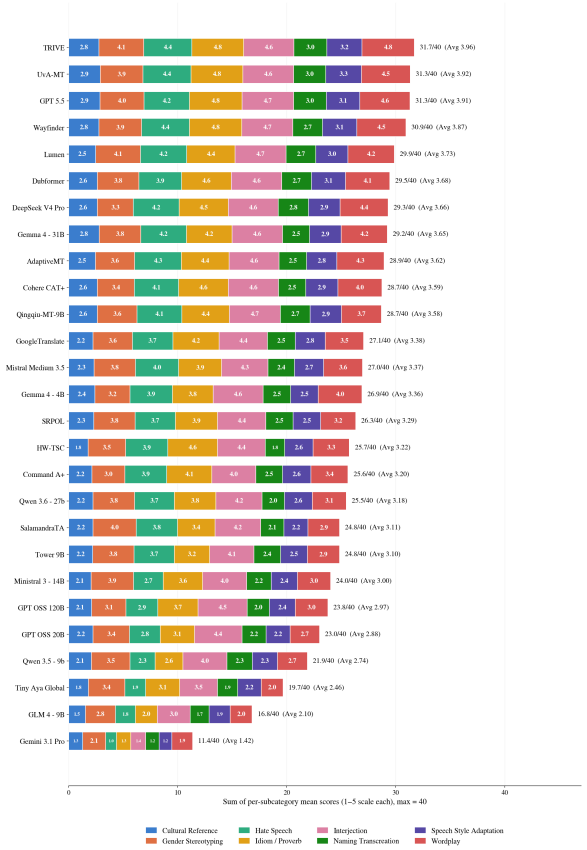


Figure 5: LLM-as-a-Judge leaderboard for Cultural Translation (CT). It displays the average scores (on a 1-5 scale) across the 8 sub-categories for 27 systems, stacked together. A larger total bar length (maximum 40) indicates superior overall performance. The systems are sorted in descending order by their overall rank, which is based on an equally weighted average across the eight sub-categories.

D KUEST Dataset Examples

D.1 Terminology & Accuracy(TA)

Table 4: Examples of Glossary Compliance examples in the Terminology & Accuracy (TA)

Doc ID	Term	Canonical Term	Source Text	Reference Text
001	electronics	전자기기	중앙처리장치. 고급 전자기기 제작에 사용된다.	(A central processor unit, useful in advanced electronics crafting.)
003	backpack	배낭	태양광 배낭	(solar backpack)
008	blade	칼날	부러진 칼날	(Broken blade)
013	blade trap	칼날 함정	칼날 함정을 2칸 떨어진 곳에 설치합니다.	(You set the blade trap 2 squares away.)
014	bow	활	활의 날개를 조여 장력을 증가시킵니다.	(You tighten the limbs, increasing the strength of the bow.)

Table 5: Examples of Polysemy in the Terminology & Accuracy (TA)

Doc ID	Term	Sense Label	Source Text	Reference Text
269	bat	박쥐	박쥐	(Bat Ears)
270	bat	방망이	당신은 야구 방망이를 휘두를 수 있습니다.	(You can swing a baseball bat.)
272	charge	충전	(%s/%s 충전량)	(%s/%s charge)
273	charge	전하	주변의 공기는 전하를 잃습니다.	(The air around you loses its charge.)
331	charge	돌격	돌격	(charge)
278	court	궁정	여름 궁정	(The Summer Court)
279	court	코트	테니스 코트	(tennis court)
280	court	법정	어떻게 하면 당신의 법정에 참여할 수 있나요?	(How can I join your court?)

D.2 Persona & Pragmatics(PP)

Table 6: Examples of Score Improvement via Regime R in the Persona & Pragmatics (PP)

Doc ID	System	Source Text	Baseline Output	Regime R Output	Score Change
071-b	AdaptiveMT	Quickly, men! Onward!	서둘러라, 제군들! 앞으로! (Quickly, men! Forward!)	신속히, 병사들이여! 전진하라! (Promptly, soldiers! Advance!)	2 → 5
083-a	GPT OSS 120B	The merfolk...they came back! Bring back the search parties! Call back the hordes!	인어들이... 놈들이 돌아왔다! 정찰단을 도로 불러들여라! 무리를 불러모아라! (The merfolk... they came back! Call the scouts back! Call the hordes!)	인어들이…… 돌아왔다! 탐색대를 당장 복귀시키고! 무리를 다시 소집해라! (The merfolk... they returned! Return the search party at once! Call the hordes back!)	1 → 5
004-b	Mistral Medium 3.5	Who is ready for a merfolk snack?	인어 식사할 친구? (Friends to eat merfolk?)	인어 요리 한 입 할 사람? (Anyone for a bite of merfolk cuisine?)	1 → 5

Table 7: Examples of over-personalization and Pragmatic Distortion in the Persona & Pragmatics (PP)

Doc ID	System	Source Text	Reference	Baseline Output	Regime R Output	Score
009-b	Command A+	Elensefar? Why him?	엘렌세파르? 왜 그 사람이지? (Elensefar? Why him?)	엘렌세파르? 왜 하필 그야? (Elensefar? Why him?)	엘렌세파르? 왜 하필 그자식이야? (Elensefar? Why that bastard?)	5 → 2
008-a	Command A+	Don' t worry about it.	걱정 말아요. (Don't worry about it.)	걱정하지 마세요. (Don't worry.)	걱정하지 마시옵소서. (Do not worry, I pray thee.)	5 → 2
048-b	AdaptiveMT	So close. So close.	아깝다. 아까워. (So close. So close.)	아깝다. 아까워. (So close. So close.)	가깝구나. 아주 가깝구나. (It is close. Very close.)	5 → 1

D.3 Technical Robustness(TR)

Table 8: Examples of Placeholder Reordering where the relative position (start, middle, end, or solo) within a segment changes between the source and reference translation in CDDA data (`evidence.reordered = true`).

Doc ID	Set	Source Text	Reference Text
001	Consistency	Your %s's %s rams into a %s with a %s	당신의 %s %s 가 %s 를 %s 로 치었다
002	Consistency	A %s %s from the %s!	%s %s 이(가) %s 에서 나온다!
011	Consistency	<npcname> bounces off a %s!	<npcname> 가 %s 를 튕겨냈습니다!
102	Consistency	You slam into a %s	당신은 %s 를 들이 받았다
002	Diverse	%d: %s; movecost %d movestr %d	디버그2 %d : %s; 이동 소모 포인트 %d 이동 체력 %d
003	Diverse	You can not activate %s! To read a description of %s, press '!', then '%c'.	%s는 작동형 바이오닉이 아니다! %s에 대한 설명을 읽으려면, '!'를 누르고 나서, '%c'를 누르라.

Table 9: Examples of Tag Preservation. All source data originates from Wesnoth. The final row is a challenging test case where emphasis on the possessive pronoun "mine" in English lacks a direct single-word equivalent in Korean, leading the reference translation to expand the emphasis span to the noun phrase "내 손자가" (19 out of 27 systems failed, §6.2).

Doc ID	Source Text	Reference Text
001	This water will make melee weapons have the arcane damage type until the end of the current scenario.	이 물은 현재 시나리오가 끝날 때까지 근접 공격에 신령 속성을 부여합니다.
003	— The Wesnoth Tactical Guide	— 웨스노스 전략 가이드
006	You can access the Recruit and Recall commands by right-clicking on a castle tile ⋯	아성에서 오른쪽 클릭을 함으로써 소집이나 재소집 명령어를 사용할 수 있습니다. ⋯
036	$player_name + newline + Share knowledge of agriculture	$player_name + 개행 + 농업에 관한 지식을 공유
050	I was wrong. You are no grandson of mine!	내가 잘못봤구나. 넌 내 손자가 아니다!

Table 10: Examples of Length Constraint.

Doc ID	Source Text	Reference Text	Allowed Length	Ref Length	Ratio
001	A soft cotton hat commonly worn by armed forces and existentialists.	부드러운 면 모자입니다. 군인과 실존주의자들이 흔히 씁니다.	68 cells	58	0.85×
004	Stylish cloak.	스타일 넘치는 멋진 망토입니다.	14 cells	30	2.14×
005	To Serve Man	사람이 먹을수 있는걸 차리기	12 cells	27	2.25×
006	sawn pipe shotgun	단축 총열 파이프 산탄총	17 cells	23	1.35×
007	light snare kit	소형 올가미 함정 세트	15 cells	21	1.40×
021	S&W 22A	S&W 22A	7 cells	7	1.00×

D.4 Cultural Transcreation(CT)

Table 11: Examples of sub-category data and system translation hypotheses with English source and reference texts.

Doc ID	Sub-category	Source Text	Reference Text	System Hypothesis
088	Speech Style Adaptation	Just who do you think you are, slug?	네놈이 뭐라도 된다고 생각하는거냐, 달팽이? (Just who do you think you are, slug?)	**Wayfinder:** 네까짓 게 도대체 뭔데, 이 굼벵이 같은 놈아? (Just who do you think you are, you slug-like jerk?)
054	Wordplay	P-p-p-pow!	프-프-프-퓹! (P-p-p-pow!)	**TRIVE:** 빠-빠-빠-빵! (B-b-b-bang!)
007	Hate Speech	Damn you, Son of the Black-Eye! You'll pay for this!	젠장맞을, 검은 눈의 아들! 댓가를 치르게 해 주마! (Damn you, Son of the Black-Eye! You'll pay for this!)	**DeepSeek V4 Pro:** 빌어먹을, 검은 눈의 아들아! 이 대가를 치르게 될 것이다! (Dammit, Son of the Black-Eye! You will pay this price!)
032	Naming Transcreation	Necrophage	시체 아귀 (Corpse ghoul)	**Qingqiu-MT-9B:** 시체식충 (Corpse insect-eater)
038	Interjection	Ouch! This hurts!	아욱! 아프다! (Ouch! It hurts!)	**AdaptiveMT:** 아야! 이거 아파! (Ouch! This hurts!)
032	Cultural Reference	The corpse is now thoroughly pulped.	그 시체는 완전히 피떡이 되었다. (The corpse has become completely bruised blood-mass.)	**DeepSeek V4 Pro:** 시체는 이제 완전히 곤죽이 되었다. (The corpse is now completely mashed pulp.)
016	Gender Stereotyping	*leans casually against the brick wall of the shelter. "Hello again, gorgeous"*	대피소의 벽돌 벽에 무심하게 기대며 "안녕하세요, 매력쟁이."라고 말했다. (*leans casually against the brick wall of the shelter. "Hello again, gorgeous"*)	**GoogleTranslate:** 쉼터의 벽돌 벽에 무심하게 기대앉으며 "또 만났네, 아름다워." (*leans casually and sits against the brick wall of the shelter. "We meet again, beautiful"*)
001	Idiom & Proverb	⋯Works best in conjunction with a high level of intelligence.	지능이 높은 캐릭터와 결합하면 금상첨화일 것입니다. (⋯It would be the icing on the cake if combined with a high-intelligence character.)	**GPT 5.5:** 높은 지능 수치와 함께할 때 가장 효과적입니다. (⋯Works most effectively when combined with high intelligence stats.)

Table 12: Examples of style flattening for sensitive expressions. These represent typical cases within the Hate Speech sub-category where multiple systems reduced the intensity of curse words or pejorative terms from the source text. Here, the "Bleach System Count" denotes the number of system outputs out of 27 evaluated as bleached by the LLM-as-a-Judge for the given item.

Doc ID	Source Text	Reference Text	System Output	Bleach	Judgment Summary
038	Our world fell apart because our leaders were as crooked as the con-men that paid for their elections.	세상이 망해버린건 윗대가리가 선거에서 표를 사는 망할 것들이기 때문이지.	GPT 5.5: 우리 세상이 무너진 건, 우리 지도자들이 자신들의 선거 자금을 대준 사기꾼들만큼이나 부패했기 때문이야. (Our world collapsed because our leaders were as corrupt as the con-men who funded their elections.)	26/27	The pejorative tone of "crooked" and "con-men" is neutralized into descriptive expressions like "corrupt leaders" and "con-men."
041	fuck no	씨발 안돼	SRPOL: 절대 안 돼 (Absolutely not)	15/27	The strong curse word "fuck" is completely omitted, leaving only a propositional negation.
028	God damn it!	빌어먹을!	GoogleTranslate: 젠장! (Damn!)	20/27	The medium-intensity curse expression is replaced with a milder interjection, lowering the emotional intensity.

E Overall Rankings and Aggregation Sensitivity

E.1 Complete System Rankings and Cross-Category Correlation

The overall rankings presented in the main text are computed based on the average rank (Mean Rank) across systems ranked 1st through 27th. Ranking ties are resolved using mean ranks. Table 13 and Table 14 show the Spearman rank correlation coefficients (ρ) across the four KUEST categories and the automated metrics.

Table 13: KUEST Inter-Category Rank Correlation (Spearman ρ, $n = 27$)

Category	TA	PP	TR	CT	KUEST Overall
TA	—	+0.714	+0.600	+0.704	+0.816
PP	+0.714	—	+0.909	+0.953	+0.970
TR	+0.600	+0.909	—	+0.861	+0.921
CT	+0.704	+0.953	+0.861	—	+0.958

Table 14: Automated Metrics vs. KUEST Rank Correlation (Spearman ρ, $n = 27$)

Metric	$\leftrightarrow$ Auto Overall	$\leftrightarrow$ KUEST Overall
BLEU	+0.710	+0.945
COMET	+0.868	+0.926
CometKiwi	+0.894	+0.449
MetricX	+0.807	+0.410
Combined (4 Metrics)	—	**+0.727**

E.2 Aggregation Sensitivity: Z-Score-Based Re-Ranking

To evaluate whether the choice of rank aggregation (Mean Rank, **MR**) introduces bias by discarding score magnitude differences, we report alternative system rankings using identical input data with swapped aggregation formulas.

- **Evaluation Blocks:** Composed of the automated metrics block (**Auto**, 16 cells), the KUEST block (**KUEST**, 4 categories), and a 1:1 equal-weighted combined block (**Combined**).
- **Aggregation Formulas:**
  - **MR (Mean Rank):** Arithmetic mean of per-category ranks (baseline criterion; lower is better).
  - **z (Standard Z-Score):** Average of standardized z-scores (higher is better).
  - **rob-z (Robust Z-Score):** Standardized using median and MAD to mitigate distortion from extreme outliers (higher is better).

E.3 Key Observations

- **High Invariance in KUEST Rankings ($\rho \geq 0.989$):** KUEST overall rankings demonstrate near-perfect consistency across all aggregation schemes. The top 5 systems ({`GPT 5.5`, `TRIVE`, `Wayfinder`, `Dubformer`, `Lumen`}) remain identical regardless of whether ranks or z-scores are used.
- **Robustness of Mean Rank Against Outliers:** Minor rank shifts observed under standard Z-scores in automated metrics are attributed to variance expansion caused by extreme length outliers in specific model outputs (e.g., verbose reasoning traces from Gemini 3.1 Pro). Applying an outlier-robust scheme (**Robust Z**) largely restores the default Mean Rank order ($\rho = 0.990$), supporting that Mean Rank is a structurally sound choice.
- **Distinct Evaluation Axes:** The correlation between Automated Metrics and KUEST remains moderate ($\rho = 0.727$) across all aggregation formulas, indicating that KUEST captures distinct evaluation axes that are unmeasured by standard metrics. Furthermore, the top 3 systems in the combined block (`GPT 5.5`, `Wayfinder`, `TRIVE`) consistently retain top positions across all formulas.

Table 15: Overall System Rankings Across Three Aggregation Schemes

System	Auto MR	Auto z	Auto rob-z	KUEST MR	KUEST z	KUEST rob-z	Combined MR	Combined z
GPT 5.5	6.94 (2)	+0.632 (1)	+0.636 (5)	1.75 (1)	+0.992 (1)	+0.961 (1)	4.34 (1)	+0.812 (1)
TRIVE	9.38 (10)	+0.524 (5)	+0.399 (11)	3.00 (2)	+0.961 (2)	+0.909 (2)	6.19 (3)	+0.742 (3)
Wayfinder	7.50 (4)	+0.573 (3)	+0.599 (6)	3.50 (3)	+0.919 (3)	+0.855 (3)	5.50 (2)	+0.746 (2)
Dubformer	10.62 (12)	+0.522 (6)	+0.289 (12)	5.50 (4)	+0.776 (5)	+0.714 (4)	8.06 (6)	+0.649 (5)
Lumen	9.00 (8)	+0.508 (7)	+0.443 (9)	6.12 (5.5)	+0.799 (4)	+0.692 (5)	7.56 (5)	+0.653 (4)
Qingqiu-MT-9B	8.69 (6)	+0.485 (8)	+0.475 (7)	6.12 (5.5)	+0.752 (6)	+0.652 (6)	7.41 (4)	+0.619 (6)
UvA-MT	13.62 (17)	+0.304 (15)	-0.048 (17)	8.00 (7)	+0.721 (7)	+0.606 (7)	10.81 (12)	+0.512 (8)
DeepSeek V4 Pro	11.00 (13)	+0.444 (10)	+0.265 (13)	8.62 (8)	+0.641 (8)	+0.496 (8)	9.81 (8)	+0.542 (7)
Gemma 4 - 31B	12.88 (15)	+0.379 (13)	+0.062 (15)	9.25 (9)	+0.603 (9)	+0.423 (9)	11.06 (13)	+0.491 (9)
AdaptiveMT	9.12 (9)	+0.444 (9)	+0.451 (8)	10.75 (10)	+0.526 (10)	+0.304 (10)	9.94 (10)	+0.485 (10)
Cohere CAT+	10.12 (11)	+0.428 (11)	+0.409 (10)	11.38 (11)	+0.451 (11)	+0.222 (11)	10.75 (11)	+0.440 (11)
GoogleTranslate	7.12 (3)	+0.547 (4)	+0.821 (2)	11.75 (12)	+0.270 (14)	+0.052 (13)	9.44 (7)	+0.409 (13)
HW-TSC	15.62 (19)	+0.140 (18)	-0.268 (20)	13.25 (13)	+0.295 (12)	-0.016 (14)	14.44 (16)	+0.218 (15)
SRPOL	6.44 (1)	+0.578 (2)	+1.145 (1)	13.38 (14)	+0.286 (13)	+0.059 (12)	9.91 (9)	+0.432 (12)
Command A+	8.62 (5)	+0.424 (12)	+0.760 (3)	15.00 (15)	+0.142 (16)	-0.152 (15)	11.81 (14)	+0.283 (14)
Mistral Medium 3.5	15.94 (20)	+0.153 (17)	-0.216 (19)	15.12 (16)	+0.129 (17)	-0.174 (17)	15.53 (18)	+0.141 (17)
Gemma 4 - 4B	13.06 (16)	+0.235 (16)	+0.007 (16)	16.00 (17)	+0.155 (15)	-0.169 (16)	14.53 (17)	+0.195 (16)
SalamandraTA	8.81 (7)	+0.378 (14)	+0.671 (4)	18.75 (18)	-0.123 (18)	-0.563 (18)	13.78 (15)	+0.128 (18)
Qwen 3.6 - 27b	15.50 (18)	+0.095 (20)	-0.208 (18)	19.12 (19)	-0.185 (19)	-0.602 (19)	17.31 (20)	-0.045 (19)
GPT OSS 120B	20.38 (21)	-0.270 (21)	-0.960 (21)	19.50 (20)	-0.235 (20)	-0.645 (20)	19.94 (21)	-0.253 (21)
Tower 9B	12.69 (14)	+0.130 (19)	+0.130 (14)	20.50 (21)	-0.343 (21)	-0.852 (21)	16.59 (19)	-0.106 (20)
GPT OSS 20B	22.38 (22)	-0.348 (22)	-1.332 (22)	22.00 (22)	-0.722 (22)	-1.238 (22)	22.19 (22)	-0.535 (22)
Gemini 3.1 Pro	26.88 (27)	-4.006 (27)	-14.404 (27)	22.12 (23)	-2.334 (27)	-3.205 (26)	24.50 (25)	-3.170 (27)
Ministral 3 - 14B	23.00 (24)	-0.625 (24)	-1.939 (24)	23.25 (24)	-0.970 (23)	-1.721 (23)	23.12 (23)	-0.797 (24)
Qwen 3.5 - 9b	22.50 (23)	-0.393 (23)	-1.588 (23)	24.25 (25)	-1.042 (24)	-1.775 (24)	23.38 (24)	-0.718 (23)
Tiny Aya Global	24.38 (25)	-0.829 (25)	-2.908 (25)	24.75 (26)	-1.380 (25)	-2.334 (25)	24.56 (26)	-1.104 (25)
GLM 4 - 9B	25.81 (26)	-1.452 (26)	-4.676 (26)	25.25 (27)	-2.083 (26)	-3.409 (27)	25.53 (27)	-1.768 (26)

Table 16: Cross-Aggregation Spearman Rank Correlation (ρ, $n = 27$)

Evaluation Block	Mean Rank ↔ Standard z	Mean Rank ↔ Standard rob-z	Standard z ↔ Robust z
Auto Metrics (16 cells)	+0.932	+0.990	+0.895
KUEST (4 categories)	+0.989	+0.994	+0.996
Combined (Equal weight)	+0.968	+0.985	—