Large Language Models for Machine Translation Quality Annotation: Humans and Models Are Both Challenged

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are considered to be a more efficient and cost-effective alternative to human judgment for Machine Translation (MT). With MT evaluation spanning a large number of language pairs, domains and levels of annotation granularity, LLMs must be thoroughly evaluated across these dimensions before being reliably used as alternatives to human evaluation. In this paper, we evaluate the performance of LLMs for two prominent MT quality evaluation schemes: Multi-Dimensional Quality Metrics (MQM) and Error Span Annotations (ESA) by comparing their agreement with human annotators. We present results on a long-context test set of 70 language pairs and the publicly available WMT23 and WMT25 data, investigating both score and error span annotation agreement across a variety of language pairs and domains. Our results show that while LLM agreement with human annotators approaches agreement between human annotators for some evaluation tasks, both vary substantially across annotation schemes, language pairs and domains and remain unreliable for most settings. Furthermore, we identify challenges facing both human and LLM annotators: humans are particularly challenged by fine-grained MQM annotations and low-resource language pairs, while LLMs struggle with minor errors, wrong language variants and error span annotation. Our results highlight the potential for improvement for both human and LLM annotation performance, possibly through human-LLM collaborative annotation pipelines that address the reliability issues identified in this work.

1 Introduction

Translation evaluation is a challenging task, which requires evaluating multiple quality dimensions simultaneously. It includes judging the faithful transfer of meaning from the source language to the target language, the naturalness and fluency of the translated text in addition to its adherence to the target language linguistic and cultural structures, and the usability of the translation in a given context. Human evaluation is the gold standard to evaluate MT models. However, human evaluation is often expensive and time-consuming, slowing down MT model development. Furthermore, human evaluation suffers from quality and reliability issues. In addition to the highly subjective nature of translation evaluation, human evaluators differ in their skills, training, motivation and leniency. Annotation tasks also differ in their design, guidelines, compensation and nature of evaluated text. All these factors contribute to unreliability of human evaluation.

Several approaches aim to improve the reliability of human evaluation. This includes aggregating multiple annotations from different annotators for the same source and translation texts. Other work relies on more fine-grained evaluations approaches that ground evaluation rating in detailed error annotations such as MQM (Lommel et al., 2014) and ESA (Kocmi et al., 2024) instead of providing a single quality rating. While these fine-grained MT evaluation approaches offer more principled way for human evaluation which improves their quality, interpretability and consistency, they are more time-consuming and expensive to obtain.

Automatic MT evaluation metrics offer a solution to approximate human evaluation in a cost and time efficient way. Early automatic metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015) rely on lexical matching between the translation and a human-produced reference translation. While being efficient to compute, they suffer from a severe limitation of penalizing correct translations which deviate from the reference wording.

With the advent of large pretrained language models, neural evaluation metrics emerged such as BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020) and MetricX (Juraska et al., 2024), offering

the ability to represent and compare text beyond the surface level, which was the main limitation of surface-based automatic MT metrics. Neural MT evaluation metrics are built by fine-tuning language models on human evaluation data, with some offering error span annotation similar to MQM such as in xCOMET (Guerreiro et al., 2023). While these model-based metrics often achieve better correlation with human judgment than surface-based metrics, they suffer from multiple limitations, mainly limited generalization ability on unseen language pairs and domains and lack of interpretability.

With LLMs showing more powerful abilities, researchers began exploring whether these models might provide a cost and time effective alternative for human evaluation, allowing for faster model iterations. LLMs have been probed for performing automatic evaluation for a range of different tasks out-of-the-box without the need for training them on human judgment data. They also offer the ability to customize the evaluation process through prompting and providing explanation. In the MT domain, Kocmi and Federmann (2023) introduced GEMBA, demonstrating that GPT-3.5 and GPT-4 prompted with simple, language-agnostic templates achieve state-of-the-art correlation with MQM human judgments on WMT22 data on the system level— without any task-specific fine-tuning.

In this work, we evaluate LLMs for MT evaluation following two popular MT evaluation schemes: MQM and ESA. MQM involves fine-grained error annotations where the quality score is automatically calculated out of error categories and severities whereas ESA provides a simpler error annotations with only severity and the score is input directly by the annotators instead of being calculated. Evaluating on these two different annotation schemes allows us to assess LLM performance across different levels of annotation granularity. We present results on a long-context test set of 70 language pairs and the publicly available WMT23 and WMT25 data, investigating both score and error span annotation agreement across a variety of language pairs and domains. We identify challenges facing both human and LLM annotators for fine-grained MT evaluation. Our analysis highlights the following main findings:

- LLM agreement with human annotators approaches agreement between human annotators for some evaluation tasks.
- Agreement between both LLM and human annotators varies substantially between annotation schemes, language pairs and domains.
- Agreement between human annotators is severely impacted when performing fine-grained error annotations in MQM and for low-resource language pairs.
- LLMs have particular challenges in identifying minor errors, flagging translations in the wrong language or variant and error span annotation, especially for rarer error categories.
- Low agreement levels highlight the need for improving both human and LLM evaluation reliability.

2 Related Work

2.1 Automatic MT Evaluation Metrics

The latest WMT25 metrics shared task (Lavie et al., 2025) included tasks for both score prediction and span-level error detection. The evaluation results showed that LLM-based metrics are the best performing metrics at the system level however they are outperformed by surface and embedding-based similarity metrics at the segment level. The task also demonstrated that span-level error detection remains the hardest challenge, with the best autoraters (all LLM-based) achieving only 9–13% character-level F1 compared to 47% for human annotators, which is consistent with our results.

Zhan et al. (2025) improved alignment of Large Reasoning Models with human judgment by post training them on synthetic data that emulates human evaluation process. They identify minor errors as being more misaligned with human annotations than major errors — consistent with our findings.

2.2 Reliability of LLMs as Annotators

Bavaresco et al. (2025) evaluated 11 LLMs on a 20-task NLP benchmark. They found that no single model consistently excels, LLMs correlate more with non-expert than expert annotations, and chain-of-thought and few-shot prompting yield inconsistent gains. They conclude that LLMs must be carefully validated before deployment as evaluators – a finding which our results also support.

Calderon et al. (2025) propose a leave-one-out statistical test for whether an LLM can replace a human annotator, finding that success is highly task and annotation aspect-dependent and that low

human inter annotator agreement (IAA) makes the test unreliable — a caveat applicable to the results in our study.

Kasner et al. (2025) evaluate LLMs on ESA error span annotation on a subset of the WMT24 human evaluation data. They found that human-human IAA consistently exceeds LLM-human IAA for span annotation, and that LLM-generated character-level span indices require post-hoc re-alignment — a challenge we independently observe and address.

2.3 AI-Assisted and Collaborative Human Annotation

Zouhar et al. (2025) showed that AI-assisted human ESA annotation increases error recall (1.6 spans per segment vs. 0.5 unassisted), improves inter-annotator consistency, and reduces annotation budget by 24% through pre-filtering of error-free segments. Riley et al. (2025) demonstrated that human revision of automatic MQM annotations (GEMBA-MQM, AutoMQM) rather than annotating from scratch surfaces missed errors, improves annotation quality without additional cost, and reduces annotation time by up to 12%.

Song et al. (2025) confirmed that performing MQM annotations for translation pairs simultaneously yields higher inter-annotator agreement on the side-by-side comparison results, better error span consistency and more reliability in identifying equal-quality translations than standard independent MQM annotations.

3 Experimental Setup

3.1 Data

We present our findings on three data sets:

- **WMT23 MQM.** The MQM annotations from the WMT23 Metrics Shared Task (Freitag et al., 2023) for two language pairs: English-to-German (en-de) and Chinese-to-English (zh-en), which include three independent annotation rounds per segment.

- **Long-Context Web MQM.** Our internal test set consisting of multi-paragraph text extracted from crawled web pages. Target translations were produced by human translators and a variety of MT systems, including neural and LLM-based models.

- **WMT25 ESA.** The Error Span Annotation (ESA) (Kocmi et al., 2024) data from the WMT25 Human Evaluation Data (Kocmi et al., 2025), covering 14 language pairs across 5 domains[1].

For our MQM annotations on the Long-Context Web MQM test set, professional annotators followed guidelines adapted from Freitag et al. (2021), selecting error spans in the source and/or target text and specifying the category and severity of each error using a specially designed interface. We conducted two annotation rounds covering all segments, yielding two independent annotations per sample from different annotators. Thus, our work presents results under two different annotation settings with different annotator pools and annotation platforms. Tables 3 and 4 show data statistics for the Long-Context Web MQM and WMT23 MQM test sets, respectively. All data sets have segments which are predominantly long-context.

3.2 Large Language Models

We use Gemini 2.5 Flash and GPT-OSS-120B for our evaluation. All inference is performed with `temperature`= 0 and `top_p`= 1. The models receive an adapted version of the annotation guidelines provided to human annotators (refer to Figure 12 for MQM and Figure 13 for ESA). For WMT23, we use the original MQM error annotation instructions from Freitag et al. (2021), which define the error taxonomy used in the WMT23 annotations.

We report results using zero-shot prompting, as adding few-shot examples did not improve LLM-human agreement. If the LLM response failed to adhere to the specified output format, we re-prompted up to five times before accepting the final response.

A critical issue emerged when calculating span annotation agreement with LLM-provided span indices: at least 80% of extracted spans based on these indices did not match the corresponding span text in the LLM output for both LLMs on MQM and ESA data sets. To correct this, we re-aligned spans by searching for the LLM-provided span text within the source (for source errors) or target (for target errors) and recording the first match start and end indices. This post-processing step was necessary to enable accurate error localization for annotation comparison.

[1] Due to access restrictions imposed by the YouTube platform, WMT25 speech domain was not included in the analysis and represents a limitation of the current work.

3.3 MQM Scoring

MQM scores are calculated out of the error annotations provided by both humans and the LLM using the standard MQM scoring formula (Freitag et al., 2021) which weighs major errors by 5, minor errors by 1 and punctuation errors by 0.1.

3.4 Annotation Evaluation

We evaluate LLM annotations across the following aspects: annotation-driven scores, error counts, and span-level annotations. We do so by comparing LLM annotation agreement with human annotations, treating each human annotation round as an independent annotator.

Score and Error Count Evaluation. We use Krippendorff's α (Krippendorff, 2004) with interval distance to measure inter-annotator agreement (IAA) on segment-level MQM and ESA scores as well as major and minor error counts per segment for MQM. Krippendorff's α is a widely-used agreement coefficient that provides several advantages over correlation measures such as Pearson and Spearman: it accounts for chance agreement, handles multiple annotations and measures agreement on absolute scores rather than merely relative co-variation — enabling assessment of how consistently annotators agree on absolute translation quality. We compute LLM-human IAA using a leave-one-out approach: each human annotation round is in turn excluded and replaced by the LLM annotation, and IAA is computed within the resulting annotator group. The reported LLM-human IAA is the average across all such groups.

Span Annotation Evaluation. We compute the following F1-based metrics to evaluate error span identification and localization:

- **Character overlap:** Measures character-level overlap between predicted and gold spans. Each predicted span is matched to the gold span with maximum character overlap (category-agnostic), and F1 is computed based on the number of overlapping characters.
- **Span overlap:** Measures F1 for predicting spans that overlap with gold spans regardless of category and severity (minimum one character overlap required).
- **Category/severity-specific span overlap:** Same as span overlap but with the requirement of having the same error category/severity when matching with gold spans.
- **No error:** F1 for correctly identifying error-free segments.

F1 metrics are computed for all human-human and human-LLM annotation pairs and averaged within each group.

4 Results

4.1 MQM

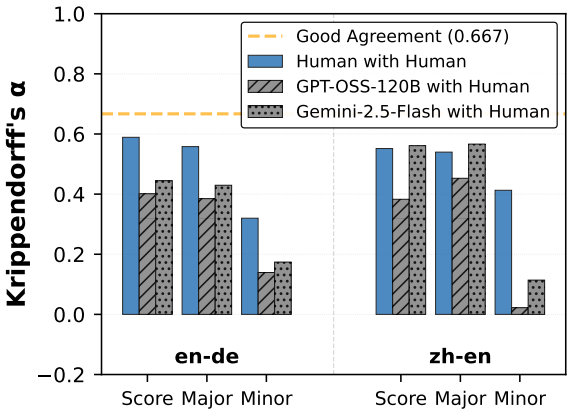


Figure 1: Inter-annotator agreement on the WMT23 MQM test set, showing Krippendorff's α for MQM scores, major error counts, and minor error counts as sub-bars within each annotator combination group.

4.1.1 Score and Error Count Evaluation

Figure 1 shows agreement for each of MQM score, major error counts and minor error counts on the WMT23 test set. For MQM score, human-human agreement is substantially better than LLM-human agreement for en-de whereas LLM-human agreement is on par with human-human agreement for zh-en. Furthermore, Gemini 2.5 Flash-human agreement exceeds GPT-OSS-120B-human agreement for both language pairs. Overall, agreement between different annotator combinations is well below the 0.667 threshold for tentative agreement. This reflects the inherent difficulty of fine-grained error annotation, particularly for long-context documents where annotator fatigue and subjective judgment become significant factors for human annotators, while LLMs may face challenges from long-context attention dispersion.

Given that MQM scores are derived from major and minor error counts, we conduct agreement experiments for segment-level major and minor error counts for different annotator pairs to better understand how each error type contributes to the overall score agreement. For major errors, LLM-human

agreement approaches or exceeds human-human agreement for zh-en but falls behind it for en-de. For minor errors, human-human agreement substantially exceeds LLM-human agreement for both language pairs, especially for zh-en where the margin is larger. This demonstrates that LLMs have more difficulty in identifying minor errors that involve subtle quality distinctions. The difficulty of minor error annotation is particularly consequential for MT evaluation: as MT system quality improves, minor errors increasingly become the primary differentiating factor among closely ranked systems, making reliable minor error annotation essential for meaningful quality estimation.

These findings are consistent with results on the Long-Context MQM test set with wider language pair coverage (Appendix 7.1), which was independently annotated using different annotator pools and annotation platform.

4.1.2 Span Annotation Evaluation

Figure 2 shows span annotation evaluation on the WMT23 test set in terms of F1 score, averaged across all language pairs. Categories are divided into high-frequency and low-frequency groups based on total segment counts. Agreement levels for high-frequency error categories are notably higher than low-frequency categories. In addition, human-human agreement exceeds LLM-human agreement for all span annotation evaluation metrics except the No Error category where Gemini 2.5 Flash-human agreement is slightly better than human-human agreement. Human-human agreement on span overlap reaches at most F1 score = 0.51, indicating the inherent difficulty of error span annotation. Notably, the agreement levels for the lower-frequency error categories are extremely low for all annotator combinations (F1 score ≤ 0.2), with many categories achieving zero F1 score by both LLMs such as Locale Convention Time Format. This suggests that while LLMs show stronger agreement with human annotators for frequent error categories such as mistranslation, they particularly struggle to identify rarer error categories such as untranslated text, wrong language and locale convention errors, possibly because these errors are underrepresented in the LLM training data. In terms of error severity, both LLM-human and human-human agreements are lower for minor errors than major errors, which confirms error count results in Section 4.1. Comparing the two LLMs, Gemini 2.5 Flash has a slightly better agreement with human annotations than GPT-OSS-120B for most error categories. Beyond error categories, agreement on identifying error-free segments is remarkably low across all annotator pairs (F1 score ≤ 0.12). This likely stems from the long-context nature of the test segments and the high disagreement on annotating minor errors: in near-perfect translations spanning multiple paragraphs, annotators frequently disagree on whether subtle imperfections constitute minor errors, leading to inconsistent error-free classifications. Span annotation results on the Long-Context MQM test set (Appendix 7.1.2) mirror this pattern with human-human agreement exceeding LLM-human for most categories, and the gap is widening for the low-frequency ones.

4.2 ESA

4.2.1 Score Evaluation

Domain	#LPs	Gemini	GPT-OSS	Both	Either
Dialogue	2	0%	50%	0%	50%
Edu	2	50%	0%	0%	50%
Literary	12	50%	16%	16%	50%
News	14	35%	21%	21%	35%
Social	14	35%	28%	21%	42%
All domains	44	38%	22%	18%	43%

Table 1: Percentage of language pairs per domain where LLM-human Krippendorff's α agreement on ESA scores exceeds human-human agreement on the WMT25 ESA test set, shown separately for each LLM, both LLMs simultaneously, and either LLM. Gemini = Gemini 2.5 Flash, GPT-OSS = GPT-OSS-120B.

Figure 3 shows ESA score agreement averaged across domains on the WMT25 ESA test set. Figures 15 and 16 in Appendix 7 show ESA score agreement for individual domains. In contrast to MQM score results (Table 2), human-human ESA score agreement exceeds LLM-human agreement for most language pairs and domains (Table 1), particularly for the news domain. This might be due to the fact that ESA has a simpler annotation scheme where scores are input directly by annotators rather than calculated from error spans, which likely reduces disagreement whereas MQM calculates the scores by weighting the span counts which might elevate the variance. However, consistent with MQM score results, ESA scores exhibit agreement below the Krippendorff's α threshold for tentative conclusions (0.667) for nearly all language pairs and domains. Exceptions include en-arz_EG across all domains, though as discussed below, the

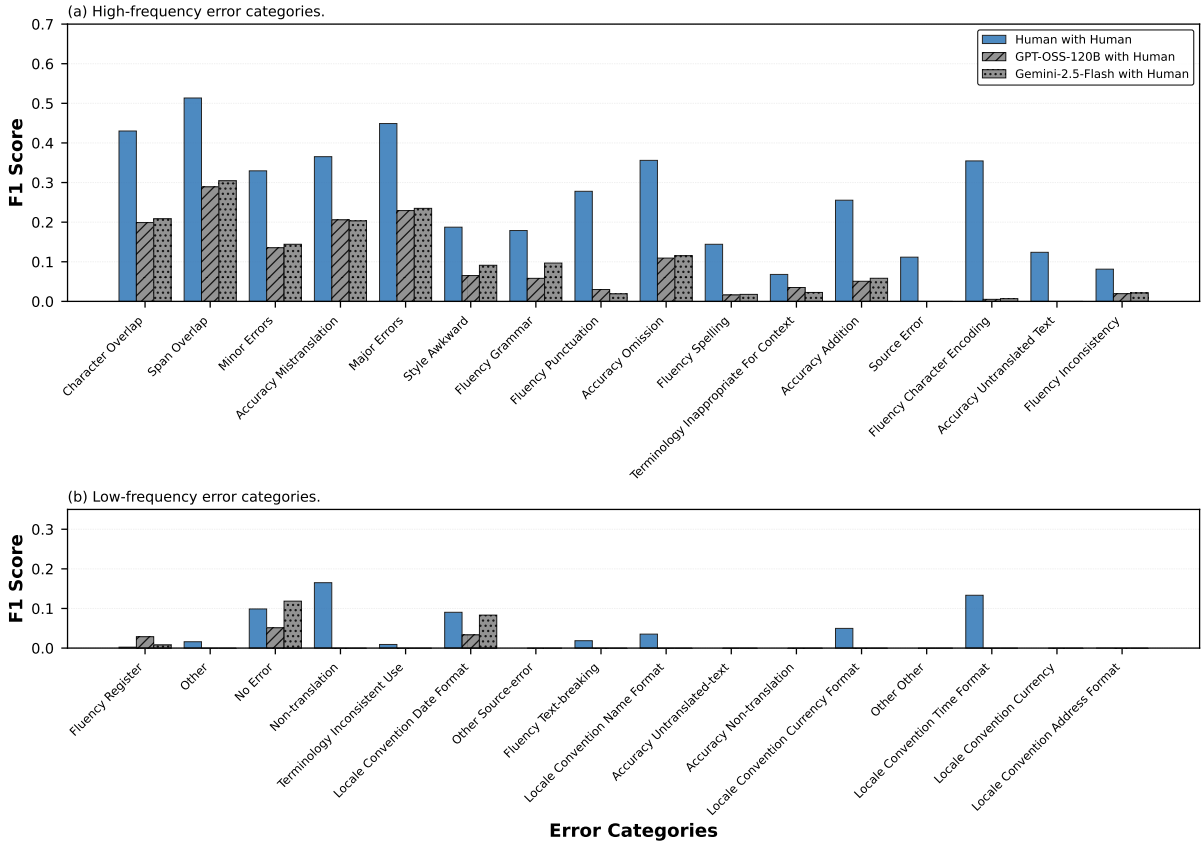


Figure 2: F1 scores for error category identification on the WMT23 MQM test set, averaged across language pairs, for (a) high-frequency and (b) low-frequency categories. Bars show agreement between annotator combinations.

high en-arz_EG agreement reflects systematic issues rather than annotation quality. Comparing the LLM-human agreement between different models, Gemini 2.5 Flash-human agreement is higher than GPT-OSS-120B-human agreement for most language pairs and domains.

Figure 4 presents ESA score agreement for each language pair and domain on the WMT25 ESA test set. Both human-human and LLM-human agreement vary substantially across domains within the same language pair, indicating that domain characteristics significantly influence annotation difficulty. For a few language pairs such as English-to-Japanese (en-ja_JP), Gemini 2.5 Flash-human agreement exceeds human-human agreement across all domains. Conversely, for language pairs such as English-to-Egyptian Arabic (en-arz_EG) and English-to-Bhojpuri (en-bho_IN), human-human agreement substantially outperforms LLM-human agreement across all domains. A notable result is English-to-Maasai (en-mas_KE), which has a remarkably low agreement across all annotator pairs consistent with its status as a low-resource language pair.

A striking finding emerges for en-arz_EG: LLM-human agreement shows negative or near zero Krippendorff's α (-0.29 for Gemini 2.5 Flash and 0.06 for GPT-OSS-120B across domains) which indicates systematic disagreement whereas human-human agreement is remarkably high (>0.90 for all domains), indicating systematic agreement. Manual examination of sample annotations reveals the root cause: Most translations (60%) for en-arz_EG are in the wrong Arabic variant (Modern Standard Arabic instead of Egyptian Arabic)[2]. While humans consistently identified these samples and scored them as zero, the LLMs consistently assigned these translations a high score because they are accurate translations despite explicit instructions in the ESA prompt (Figure 13) to score translation in the wrong language/variant by zero. This failure mode highlights a critical limitation in the LLMs' ability to distinguish between closely re-

[2]Percentage of en-arz_EG translations in wrong variant are estimated by calculating percentage of translations which has a score of 0 by one of the human annotators and a single error span, which is how annotators are instructed to treat translations with wrong language according to the ESA annotation instructions in Figure 13

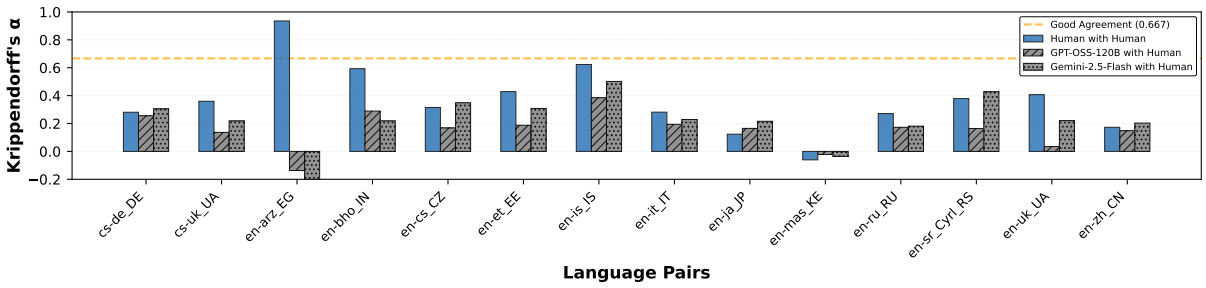


Figure 3: Inter-annotator agreement for ESA scores on the WMT25 ESA test set, averaged across all domains. Bars show Krippendorff's α for each annotator combination.

lated language variants—a capability essential for locale-specific quality evaluation.

4.2.2 Span Annotation Evaluation

Figure 5 shows span annotation evaluation on the WMT25 ESA test set in terms of F1 score, averaged across all language pairs and domains. Unlike MQM, ESA uses a smaller set of error categories, so all categories are shown in a single figure. Similar to the MQM span annotation results in Section 4.1.2, human-human agreement is substantially better than LLM-human agreement for all span annotation evaluation metrics. Notably, both LLMs struggle with identifying missing translations compared to humans. Comparing different LLMs' agreement with human annotations, Gemini 2.5 Flash and GPT-OSS-120B exhibit similar agreement levels for all error categories except for No Error category where Gemini 2.5 Flash's agreement surpasses that of GPT-OSS-120B. In terms of agreement for the two severity levels, LLM-human agreement on major errors is higher than that of minor errors, consistent with MQM results in Section 4.1.2.

4.3 Annotation Quality Validation

We observe that for many language pairs, there is a substantial difference in the LLM-human agreement across human annotation rounds. For example Figure 6 shows that Gemini 2.5 Flash's agreement with each human annotation round on MQM scores for English-to-Hindi (en_US-hi_IN), English-to-Vietnamese (en_US-vi_VN) and English-to-Thai (en_US-th_TH). For these language pairs the human-human agreement is negative or close to zero and Gemini 2.5 Flash's agreement with one of the human annotation rounds substantially exceeds its agreement with the other. This suggests that the two human annotation rounds are themselves inconsistent for these language pairs and that LLMs can serve as diagnostic tools for flagging human annotation inconsistencies for further investigation.

5 Conclusion

We evaluate LLMs as MQM and ESA annotators using inter-annotator agreement on a long-context test set of 70 language pairs and the publicly available WMT23 and WMT25 data, investigating both score and error span annotation agreement across a variety of language pairs, domains and annotation settings, using two LLMs: Gemini 2.5 Flash and GPT-OSS-120B. We demonstrate that LLM-human agreement exceeds human-human agreement for MQM scores for the majority of MQM language pairs and a subset of ESA language pairs and domains, which highlights the potential of using LLMs for machine translation quality estimation. However, our findings suggest that LLMs are not yet a reliable substitute for human MT evaluation. First, LLM-human agreement levels vary considerably between different language pairs and domains and between different LLMs. This highlights the need for conducting thorough LLM-human agreement tests with tight quality controls before using LLMs for MT evaluation in specific domains and language pairs. Second, agreement levels remain well below the Krippendorff's α threshold for tentative conclusion agreement (0.667), particularly for low-resource language pairs, highlighting a substantial potential for enhancing both human and LLM annotation performance to achieve more reliable annotations.

Our results identify several key challenges facing both humans and LLMs for machine translation quality evaluation. LLM annotators have particularly low agreement with human annotators on minor errors. Furthermore, LLMs struggle with error span annotation for all error categories, particularly

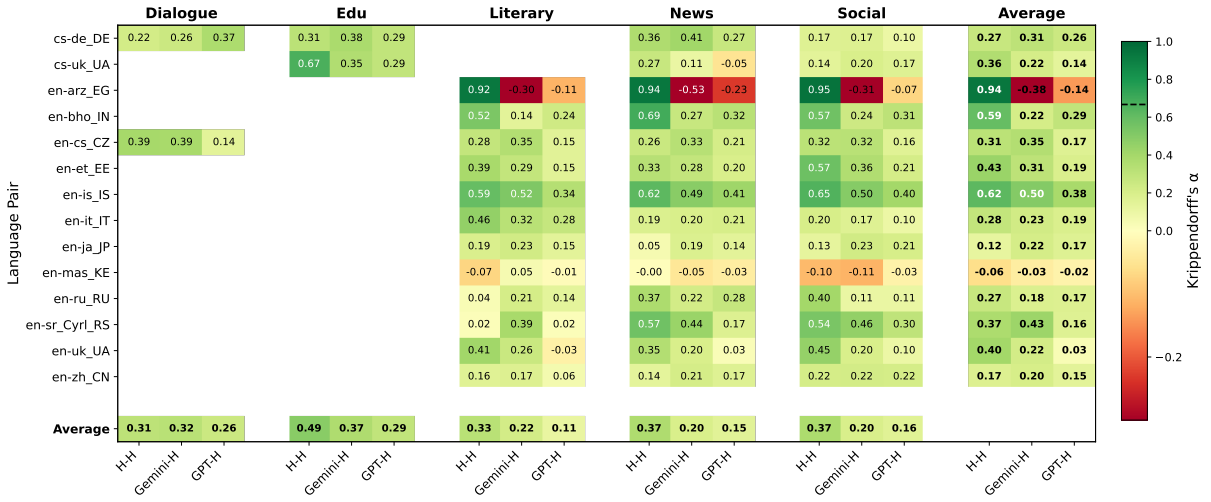


Figure 4: Krippendorff's α agreement on ESA scores for each language pair and domain on the WMT25 ESA test set. Within each domain, columns show H-H (human-human), Gemini-H (Gemini 2.5 Flash vs human), and GPT-H (GPT-OSS-120B vs human) agreement. The rightmost group and bottom row show averages across domains and language pairs respectively. The dashed line on the colorbar marks $\alpha = 0.667$.

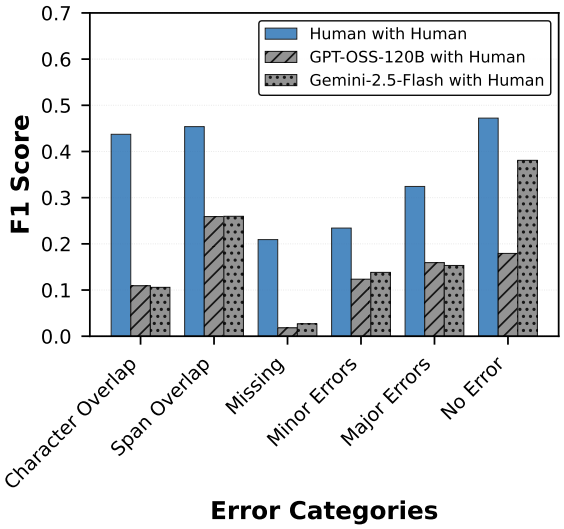


Figure 5: F1 scores by error category on the WMT25 ESA test set. Bars show agreement between annotator combinations.

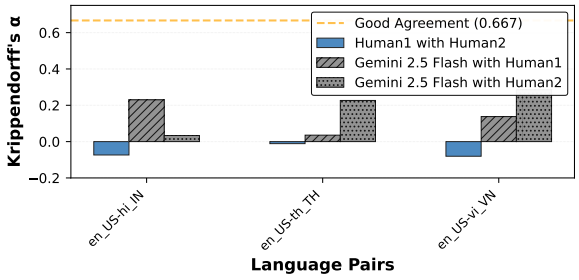


Figure 6: Inter-annotator agreement for MQM scores on the Long-Context Web MQM test set for en_US-hi_IN, en_US-vi_VN, and en_US-th_TH, showing Gemini 2.5 Flash's agreement with each individual human annotation round and agreement between human annotation rounds.

translations in the wrong language variant, untranslated text and rarer locale-specific errors, suggesting that targeted prompt engineering or fine-tuning may offer a path to improvement. Human annotators are particularly challenged on fine-grained MQM error annotation and low-resource language pairs, which needs to be addressed through clearer guidelines, better annotator training and improved annotation processes that reduce annotator fatigue. We also show that LLMs may be used as diagnostic tools to flag inconsistencies in human annotations. Although the identified weaknesses could be addressed independently for LLMs and humans, we believe a combined LLM-human collaborative framework that leverages both complementary strengths would offer a more promising path to reliable and efficient MT evaluation.

6 Limitations

One limitation of this work is that the Krippendorff's α threshold for reliable agreement is not well-established for the fine-grained MT quality annotations such as MQM. The current widely used 0.667 threshold in the research community for establishing tentative agreement might not be realistic for such a task with inherent difficulty and subjectivity. More research is needed to establish realistic annotator reliability thresholds under different set-

tings such as text length, domain and number of annotations.

For licensing reasons we cannot publish our internally-annotated Long-Context Web MQM test sets, which presents a significant limitation of the replicability of the MQM results presented in this work. While we tried to cover a variety of language pairs, domains and annotation schemes in our work, more studies should be conducted to gauge the LLM ability for MT quality evaluation on larger data sets spanning a wider set of language pairs and domains.

We presented our results on only two LLMs: Gemini 2.5 Flash and GPT-OSS-120B using a single prompt in a zero-shot setting with a simple heuristic for span correction. With the growing variety of available LLMs, more evaluation needs to be conducted using different LLMs with ablation for different prompting strategies and a better approach for span localization.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. *Preprint*, arXiv:2501.10970.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Juraj Juraska, Daniel Deutsch, Mara 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*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. Large language models as span annotators. *Preprint*, arXiv:2504.08697.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 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*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

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*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Parker Riley, Daniel Deutsch, Mara Finkelstein, Colten DiIanni, Juraj Juraska, and Markus Freitag. 2025.

MQM re-annotation: A technique for collaborative evaluation of machine translation. *Preprint*, arXiv:2510.24664.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing human evaluation in machine translation with comparative judgment. *Preprint*, arXiv:2502.17797.

Runzhe Zhan, Zhihong Huang, Xinyi Yang, Lidia S. Chao, Min Yang, and Derek F. Wong. 2025. Are large reasoning models good translation evaluators? Analysis and performance boost. *Preprint*, arXiv:2510.20780.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

7 Appendix

7.1 Results on Long-Context Web MQM Test Set

7.1.1 Score and Error Count Evaluation

Figure 7a shows MQM score agreement averaged by language pair direction. For en-X and X-Y language pairs, LLM-human average agreement approaches or exceeds human-human average agreement. However, for X-en language pairs, human-human average agreement slightly outperforms LLM-human average agreement. LLMs' performance differed by language pair direction too. Gemini 2.5 Flash agreement was overall better than GPT-OSS-120Bfor X-Y language pairs and en-X language pairs, while the opposite is true for X-en language pairs.

Figures 7b and 7c show average agreement by language pair direction for major and minor error counts, respectively. Human-human agreement on major error counts is higher than minor error agreement for X-en and X-Y language pairs while the opposite is true for en-X. Similar to the WMT23 test set results in Section 4.1, both LLMs show substantially lower agreement for minor errors counts than major error counts.

Direction	#LPs	Gemini	GPT-OSS	Both	Either
en-X	19	57%	42%	36%	63%
X-en	19	47%	47%	36%	57%
X-Y	32	84%	78%	75%	87%
All	70	67%	60%	54%	72%

Table 2: Percentage of language pairs per translation direction where LLM-human Krippendorff's α on MQM scores exceeds human-human agreement on the Long-Context Web MQM test set. Gemini = Gemini 2.5 Flash, GPT-OSS = GPT-OSS-120B.

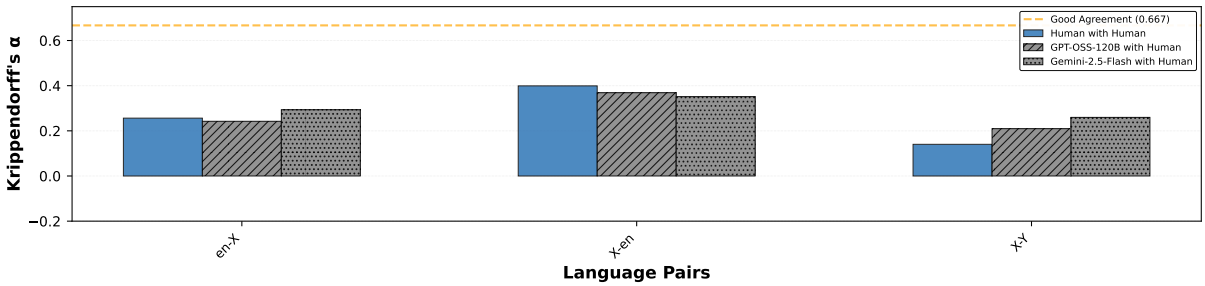
Table 2 shows the percentage of language pairs where LLM-human score agreement exceeds human-human score agreement. Comparing LLM-human agreement to human-human agreement, LLM-human agreement exceeds human-human agreement for most language pairs (67% for Gemini 2.5 Flash, 60% for GPT-OSS-120B), particularly for X-Y direction (84% for Gemini 2.5 Flash and 78% for GPT-OSS-120B) while the opposite is true for X-en language pairs where human-human agreement exceeds LLM-human agreement for 53% of language pairs.

Figures 9–11 present MQM score agreement results for individual language pairs within each direction. Both LLM-human and human-human agreement vary substantially between different language pairs. However, the overall agreement falls well below the Krippendorff's α = 0.667 threshold for tentative conclusions for nearly all language pairs, with rare exceptions where human-human agreement reaches this threshold, including Russian-to-English (ru_RU-en_US), Simplified Chinese-to-English (zh_CN-en_US) and Traditional Chinese-to-English (zh_TW-en_US). Agreement is particularly low for Non-English pairs, which predominantly comprise low-resource languages. For humans this might indicate issues in the expertise and/or training of the human annotators for these language pairs. For LLMs, this is expected as lower-resource language pairs would be underrepresented in the LLM training data.

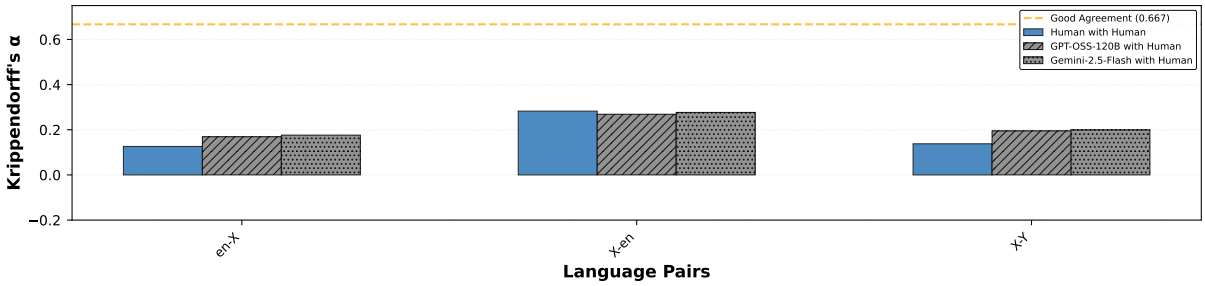
7.1.2 Span Annotation Evaluation

Figure 8 presents F1 scores for error category identification across annotator pairs, averaged over all language pairs. Similar to the WMT23 test set results in Section 4.1.2, agreement levels are low across all categories and annotator pairs with agreement reaching at most F1 score = 0.36.

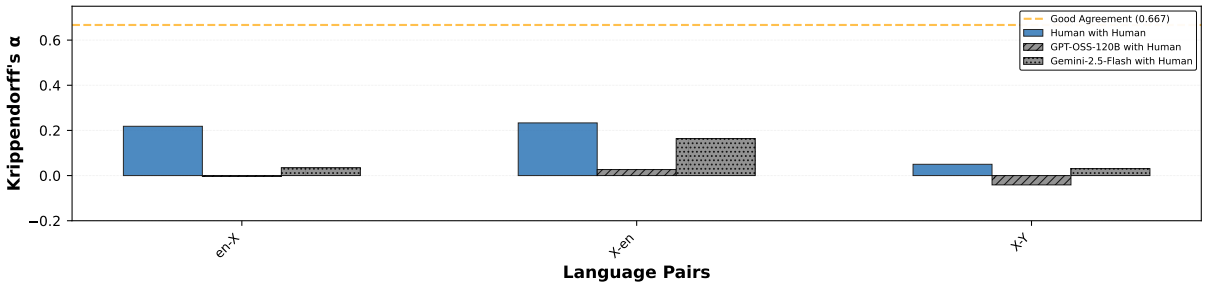
Human-human agreement exceeds LLM-human agreement for all span evaluation metrics except for mistranslation, major errors and non-translation where Gemini 2.5 Flash-human agreement matches or exceeds human-human agreement. GPT-OSS-120B-human agreement is lower than Gemini 2.5 Flash-human agreement for most categories. Both LLMs achieve low F1 for several low-frequency categories, particularly Locale Convention errors (F1 score < 0.12) and have notably low agreement on identifying error-free segments across all annotator pairs (F1 score < 0.18) consistent with the WMT23 test set results in Section 4.1.2.



(a) MQM scores.



(b) Major error counts.



(c) Minor error counts.

Figure 7: Inter-annotator agreement by translation direction on the Long-Context Web MQM test set, showing mean Krippendorff's α averaged across all language pairs in each direction group (en-X, X-en, X-Y), for (a) MQM scores, (b) major error counts, and (c) minor error counts.

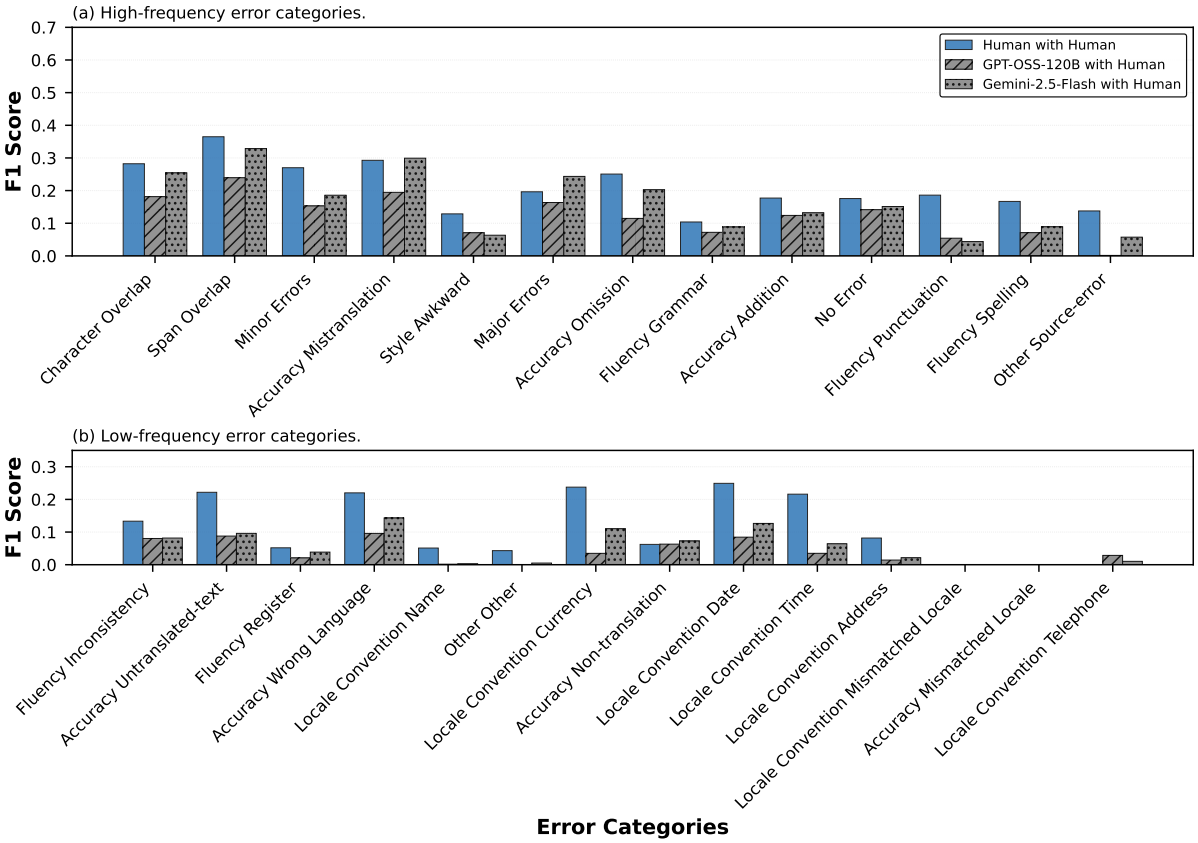
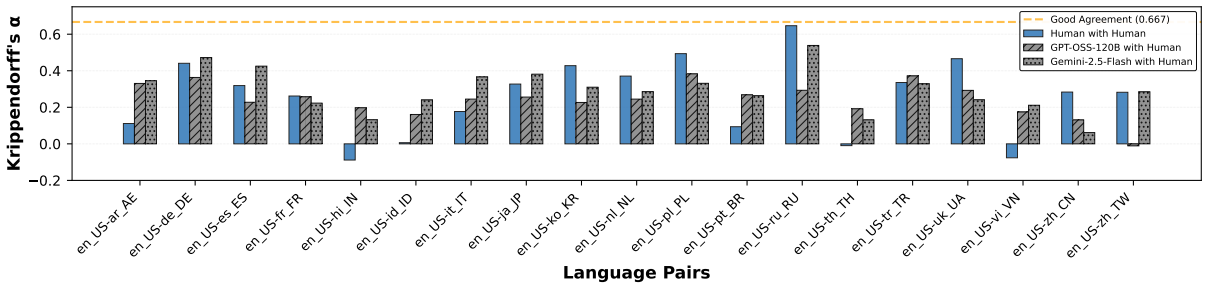
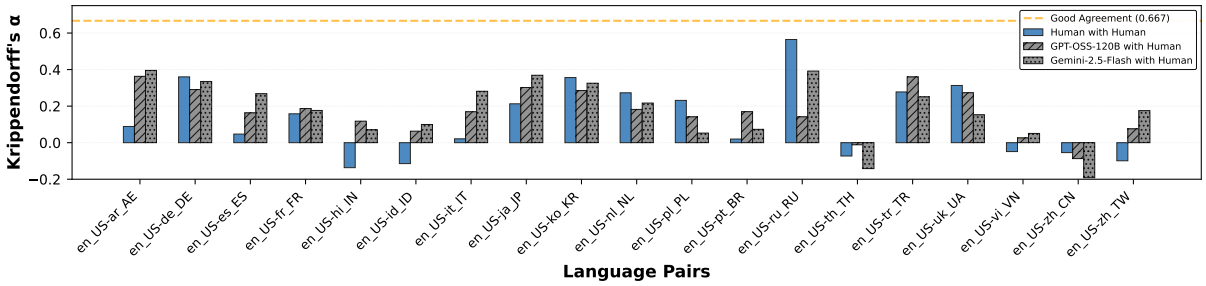


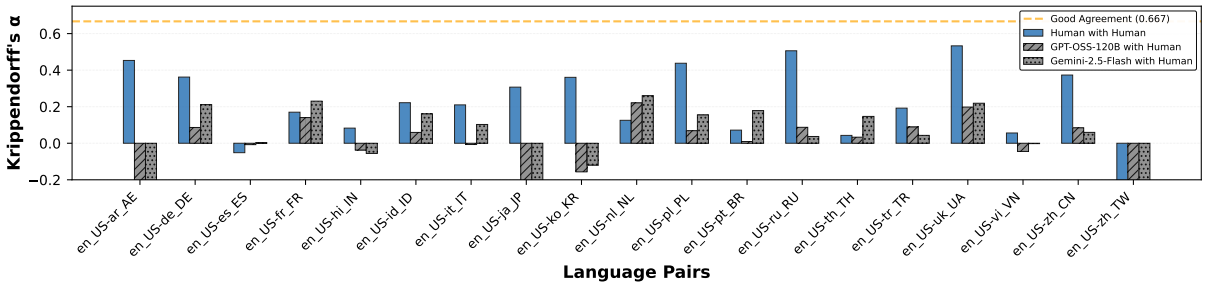
Figure 8: F1 scores for error category identification on the Long-Context Web MQM test set, averaged across language pairs, for (a) high-frequency and (b) low-frequency categories. Bars show agreement between annotator combinations.



(a) MQM scores.

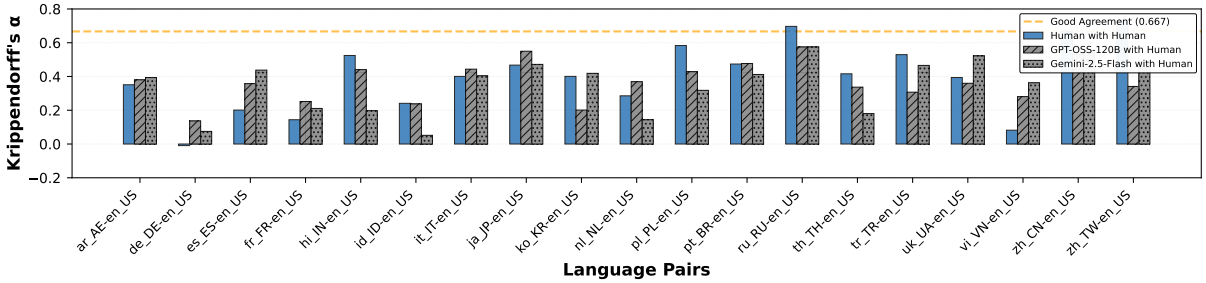


(b) Major error counts.

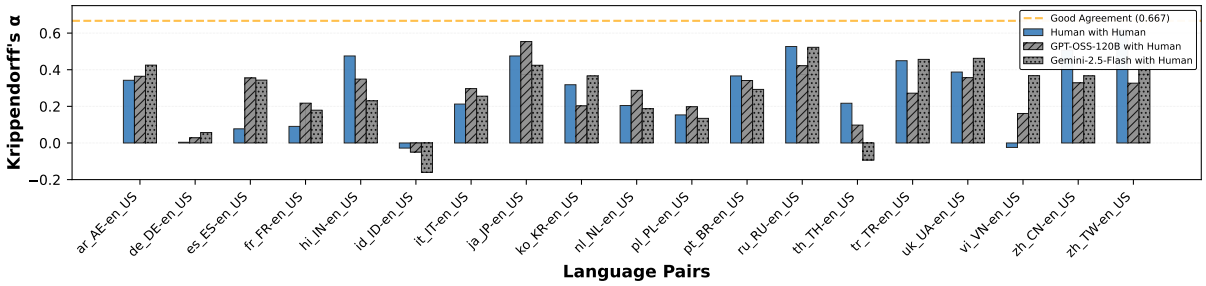


(c) Minor error counts.

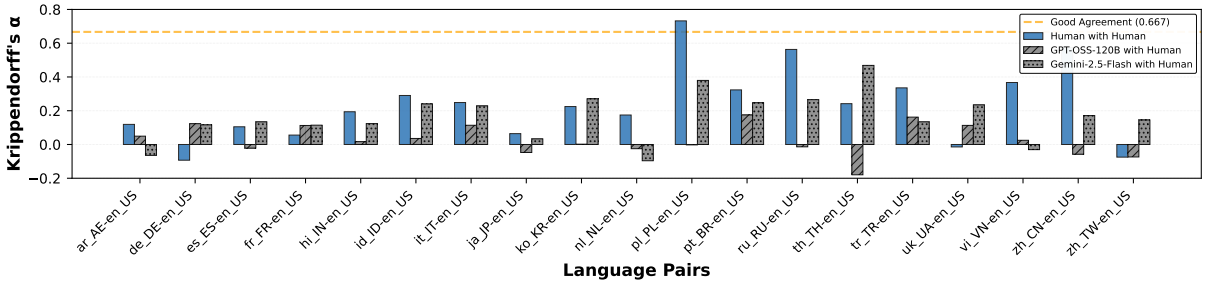
Figure 9: Inter-annotator agreement on the Long-Context Web MQM test set for English→Other (en-X) language pairs, for (a) MQM scores, (b) major error counts, and (c) minor error counts. Bars show Krippendorff's α for each annotator combination.



(a) MQM scores.

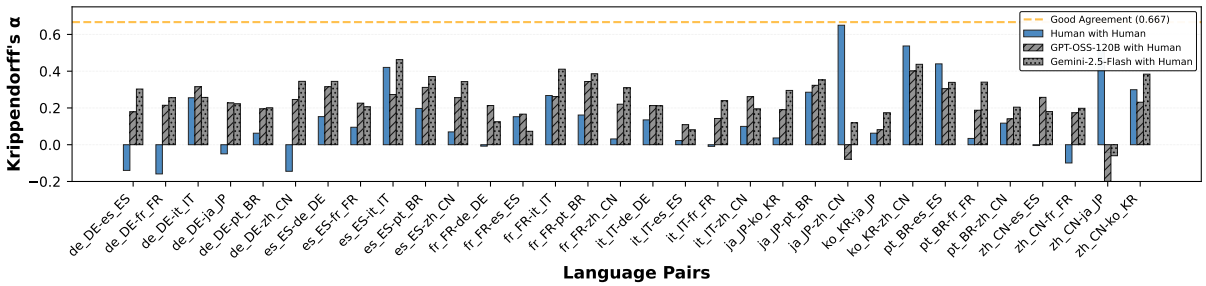


(b) Major error counts.

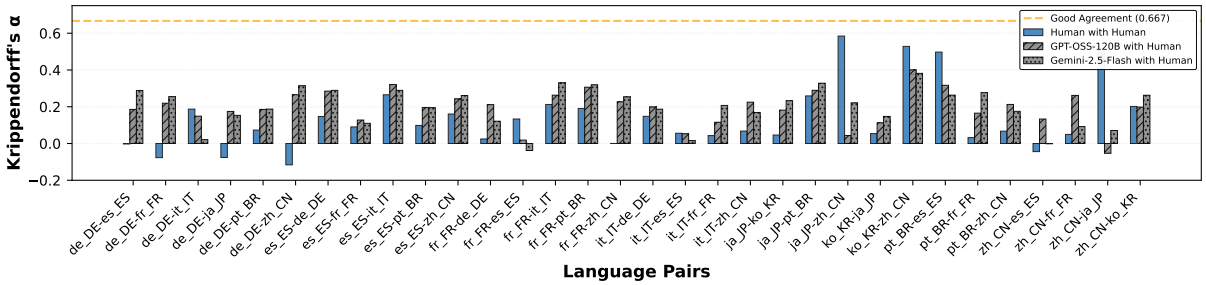


(c) Minor error counts.

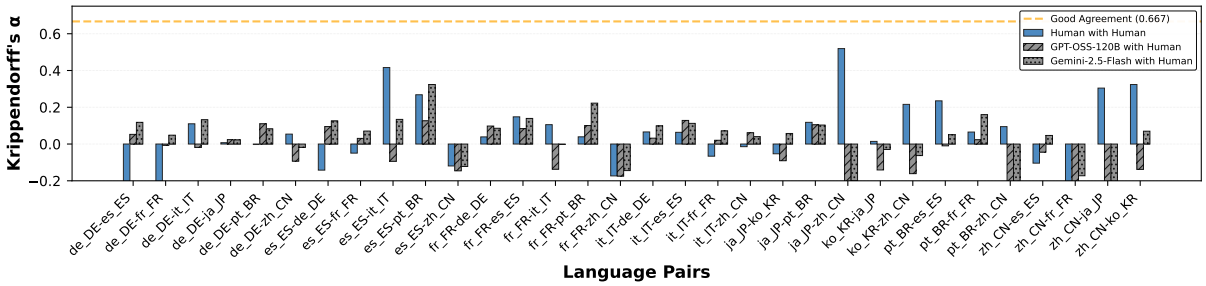
Figure 10: Inter-annotator agreement on the Long-Context Web MQM test set for Other→English (X-en) language pairs, for (a) MQM scores, (b) major error counts, and (c) minor error counts. Bars show Krippendorff's α for each annotator combination.



(a) MQM scores.



(b) Major error counts.



(c) Minor error counts.

Figure 11: Inter-annotator agreement on the Long-Context Web MQM test set for Other→Other (X-Y) language pairs, for (a) MQM scores, (b) major error counts, and (c) minor error counts. Bars show Krippendorff's α for each annotator combination.

```
Translation Quality Evaluation

## Task Description
You are an expert translation quality evaluator.
You are given a text in the source language and a corresponding translation in the target language.
Your task is to evaluate the translation using Multidimensional Quality Metrics (MQM) according to
     instructions defined below.

## MQM Error Annotation Instructions
Please follow these guidelines to annotate all translation errors.

Please read the source and translation texts and annotate translation errors one by one, taking the context
     of the whole text into consideration. For each translation error please provide the following
     information:
Span: select the error span(s)
Category: select the error category from the error categories defined below.
Severity: select the error severity from the error severities defined below.

### Error Severity Levels
Severity    Description
Major       Errors that may confuse or mislead the reader due to significant change in meaning.
Minor       Errors that don't lead to loss of meaning and wouldn't confuse or mislead the reader but would
     be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less
     appealing.


### Error Categories
Error Category                  Description
Accuracy/Addition               Translation includes information not present in the source
Accuracy/Omission               Translation is missing content from the source
Accuracy/Mistranslation         Translation does not accurately represent the source
Accuracy/Wrong Language         Translation includes segments in wrong target language. This doesn't include
      scenarios like names or quotes that are intended to be in another language.
Accuracy/Mismatched Locale      Translation includes segments in wrong locale variant.
Accuracy/Untranslated-text      Source text was left untranslated in the target when a translation is
     expected
Accuracy/Non-translation        Impossible to reliably characterize distinct translation errors because the
     translation is garbled or is unrelated to the source text
Fluency/Punctuation             Incorrect punctuation for locale or style
Fluency/Spelling                Incorrect spelling or capitalization
Fluency/Grammar                 Problems with grammar, other than orthography
Fluency/Register                Wrong grammatical register (eg, inappropriately informal pronouns).
Fluency/Inconsistency           Translation of the same word/phrase is inconsistent across the text
Style/Awkward                   Translation has inappropriate language style despite being accurate and
     grammatical
Locale convention/Address       Address format does not follow target locale conventions
Locale convention/Currency      Currency format does not follow target locale conventions
Locale convention/Date          Date format does not follow target locale conventions
Locale convention/Name          Name format does not follow target locale conventions
Locale convention/Telephone     Telephone number format does not follow target locale conventions
Locale convention/Time          Time format does not follow target locale conventions
Other/Source-error              An error in the source. (e.g. the text is not in the expected language, is
     garbled, nonsensical, has spelling errors or typos or non coherent)
Other/Other                     Any other issue in the source or translation

## Output Format
Return a JSON object with the following field:
- errors (list): list of translation errors found. Each error is an object with:
  - category (string): error category from the predefined list above (e.g. "Accuracy/Mistranslation")
  - severity (boolean): true for Major errors, false for Minor errors
  - other (string, optional): free-text description when category is "Other/Other"
  - spans (list): one or more error spans, each with:
    - start (integer): character index of the first character of the span (0-based, relative to the
         paragraph)
    - end (integer): character index of the last character of the span (inclusive, relative to the paragraph
         )
    - paraIndex (integer): index of the paragraph containing the span (0-based)
    - spanText (string): the actual text of the selected span
    - isSource (boolean): true if the span is on the source text, false if on the target
If there are no errors, return an empty list: {"errors": []}

## Translation to Evaluate
Source (<source_language>): <source_text>
Translation (<target_language>): <target_text>

## Your Evaluation
```

Figure 12: MQM LLM evaluation prompt template.

```
Translation Quality Evaluation

## Task Description
You are an expert translation quality evaluator.
You are given a text in the source language and a corresponding translation in the target language.
Your task is to evaluate the translation using Error Span Annotation according to instructions defined below
     .

## Error Annotation Instructions
Select the part of translation where you have identified a translation error. Choose error severity:
- Minor errors: Style, grammar, word choice could be better or more natural.
- Major errors: The meaning is changed significantly and/or the part is really hard to understand.

Tips:
- Omitted content: If the translation is missing content from the source, set both start_i and end_i to "
     missing".
- Highlight the word or general area of the error (it doesn't need to be exact). Use multiple highlights for
      different errors.
- Pay particular attention to the language/variant being evaluated. If the translation is in the wrong
     language or language variant, mark it fully and assign it 0.
- Pay particular attention to translation consistency between texts across the whole document.
- If the translation contains additional text (e.g. "Here is the translation") or alternative secondary
     translation, mark it as a major error.

Rating Scale (from best to worst):

- Perfect [95-100]: The meaning of the translation is completely consistent with the source and reads like
     it was formulated by a native speaker.
- Retains core meaning with minor mistakes [67-94]: The translated sentence retains the core meaning of the
     source sentence, but may have some minor grammar mistakes, awkward formulations and/or might omit some
     insignificant parts.
- Preserves some meaning with major mistakes [6-66]: The translated sentence preserves some of the meaning
     of the source sentence but misses significant parts, or contains major errors.
- Nonsense or unrelated [0-5]: Completely nonsense translation that is not related to the source string.

Within each rating category, better translations should get higher scores.

## Output Format
Return a JSON object with the following fields:
- score (float): overall quality score in the range [0-100]
- errors (list): list of translation errors found. Each error is an object with:
  - start_i (integer or "missing"): character index of the first character of the error span (0-based), or "
       missing" if the error has no localized span (e.g. omission)
  - end_i (integer or "missing"): character index of the last character of the error span (inclusive), or "
       missing" if the error has no localized span
  - spanText (string): the actual text of the selected span, or null if missing
  - severity (boolean): true for Major errors, false for Minor errors
If there are no errors, return an empty list: {"score": <score>, "errors": []}

## Translation to Evaluate
Source (<source_language>): <source_text>
Translation (<target_language>): <target_text>

## Your Evaluation
```

Figure 13: ESA LLM evaluation prompt template adapted from ESA guidelines (Kocmi et al., 2024)

```
{
  "srcLocale": "es_ES",
  "tgtLocale": "en_US",
  "id": 200,
  "annotatorID": 5933470,
  "source": "Acceso Cliente Si aun no esta registrado, registrese aqui (8 caracteres y debe incluir numeros,
       letras y algun caracter especial) Si no puede acceder a su cuenta, solicite una nueva contrasena
      aqui Configuracion de cookies del sitio web A continuacion puedes permitir o bloquear las cookies.
      Recuerda en caso de que no permitas la instalacion de cookies en tu navegador es posible que tu
      experiencia en nuestra web pueda resultar menos satisfactoria. Puede conocer la informacion completa
      sobre el uso de las cookies, su configuracion, origen, finalidades y derechos en nuestra Politica de
      cookies. Usted permite el uso de las cookies para las siguientes finalidades:",
  "target": "Client Access If you are not registered yet, register here (8 characters and must include
      numbers, letters and some special character) If you cannot access your account, request a new
      password here Website cookie settings Then you can allow or block cookies. Remember that if you do
      not allow the installation of cookies in your browser, it is possible that your experience on our
      website may be less satisfactory. You can know the complete information about the use of cookies,
      their configuration, origin, purposes and rights in our Cookies Policy. You allow the use of cookies
      for the following purposes:",
  "skip": false,
  "skip_reason": null,
  "non-reviewable": false,
  "non-reviewable_reason": null,
  "errors": [
    {
      "category": "Accuracy/Mistranslation",
      "severity": false,
      "other": null,
      "spans": [
        {
          "start": 222,
          "end": 225,
          "paraIndex": 0,
          "spanText": "Then",
          "isSource": false,
        },
        {
          "start": 237,
          "end": 250,
          "paraIndex": 0,
          "spanText": "A continuacion",
          "isSource": true,
        }
      ]
    },
    {
      "category": "Fluency/Spelling",
      "severity": false,
      "other": null,
      "spans": [
        {
          "start": 198,
          "end": 220,
          "paraIndex": 0,
          "spanText": "Website cookie settings",
          "isSource": false,
        }
      ]
    },
    {
      "category": "Style/Awkward",
      "severity": false,
      "other": null,
      "spans": [
        {
          "start": 415,
          "end": 426,
          "paraIndex": 0,
          "spanText": "You can know",
          "isSource": false,
        }
      ]
    }
  ]
}
```

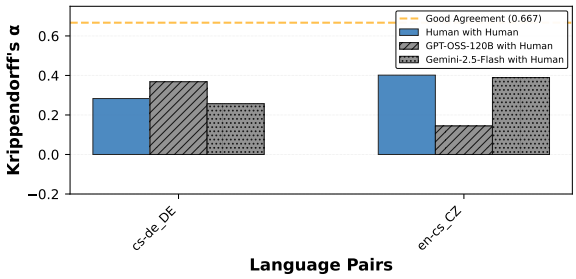

Figure 14: Example of MQM annotation in JSON format showing translation from Spanish to English with error annotations including category, severity, and character-level span positions.

Table 3: Statistics for Long-Context MQM Test Sets. Len shows average source segment length in words (^{w}) for space-using languages or characters (^{c}) for non-space languages.

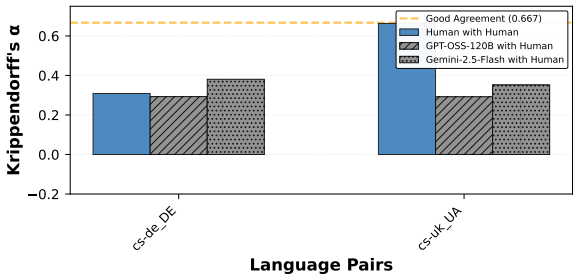
LP	Segs	Sys	Ann	Len	LP	Segs	Sys	Ann	Len
ar_AE-en_US	100	1	11	136^{w}	fr_FR-es_ES	200	4	6	157^{w}
de_DE-en_US	200	4	10	136^{w}	fr_FR-it_IT	200	4	11	157^{w}
de_DE-es_ES	200	4	13	136^{w}	fr_FR-pt_BR	200	4	19	157^{w}
de_DE-fr_FR	200	4	23	136^{w}	fr_FR-zh_CN	200	4	15	157^{w}
de_DE-it_IT	200	4	13	136^{w}	hi_IN-en_US	100	1	11	142^{w}
de_DE-ja_JP	200	4	16	136^{w}	id_ID-en_US	100	1	10	133^{w}
de_DE-pt_BR	200	4	23	136^{w}	it_IT-de_DE	200	4	18	142^{w}
de_DE-zh_CN	200	4	19	136^{w}	it_IT-en_US	200	4	12	140^{w}
en_US-ar_AE	100	1	8	138^{w}	it_IT-es_ES	200	4	12	142^{w}
en_US-de_DE	200	3	19	136^{w}	it_IT-fr_FR	200	4	16	142^{w}
en_US-es_ES	200	3	18	136^{w}	it_IT-zh_CN	200	4	18	142^{w}
en_US-fr_FR	200	3	17	136^{w}	ja_JP-en_US	200	4	9	195^{c}
en_US-hi_IN	100	1	11	138^{w}	ja_JP-ko_KR	200	4	13	390^{c}
en_US-id_ID	100	1	7	138^{w}	ja_JP-pt_BR	200	4	15	390^{c}
en_US-it_IT	200	3	22	136^{w}	ja_JP-zh_CN	200	4	9	390^{c}
en_US-ja_JP	200	3	10	136^{w}	ko_KR-en_US	200	4	10	141^{w}
en_US-ko_KR	200	3	5	136^{w}	ko_KR-ja_JP	200	4	17	99^{w}
en_US-nl_NL	100	1	6	138^{w}	ko_KR-zh_CN	200	4	18	99^{w}
en_US-pl_PL	100	1	6	138^{w}	nl_NL-en_US	100	1	5	137^{w}
en_US-pt_BR	200	3	19	136^{w}	pl_PL-en_US	100	1	5	141^{w}
en_US-ru_RU	100	1	4	138^{w}	pt_BR-en_US	200	4	19	141^{w}
en_US-th_TH	100	1	3	138^{w}	pt_BR-es_ES	200	4	17	146^{w}
en_US-tr_TR	100	1	8	138^{w}	pt_BR-fr_FR	200	4	15	146^{w}
en_US-uk_UA	100	1	5	138^{w}	pt_BR-zh_CN	200	4	19	146^{w}
en_US-vi_VN	100	1	6	138^{w}	ru_RU-en_US	100	1	6	139^{w}
en_US-zh_CN	200	3	8	136^{w}	th_TH-en_US	50	1	3	884^{c}
en_US-zh_TW	100	1	10	138^{w}	tr_TR-en_US	100	1	8	138^{w}
es_ES-de_DE	200	4	18	150^{w}	uk_UA-en_US	100	1	9	137^{w}
es_ES-en_US	200	4	21	136^{w}	vi_VN-en_US	100	1	5	144^{w}
es_ES-fr_FR	200	4	18	150^{w}	zh_CN-en_US	200	4	8	215^{c}
es_ES-it_IT	200	4	21	150^{w}	zh_CN-es_ES	200	4	7	276^{c}
es_ES-pt_BR	200	4	20	150^{w}	zh_CN-fr_FR	200	4	17	276^{c}
es_ES-zh_CN	200	4	8	150^{w}	zh_CN-ja_JP	200	4	7	276^{c}
fr_FR-de_DE	200	4	26	157^{w}	zh_CN-ko_KR	200	4	13	276^{c}
fr_FR-en_US	200	4	17	138^{w}	zh_TW-en_US	100	1	6	202^{c}

Table 4: Statistics for the WMT23 MQM test set. Len shows average source segment length in words (^{w}) for space-using languages or characters (^{c}) for non-space languages.

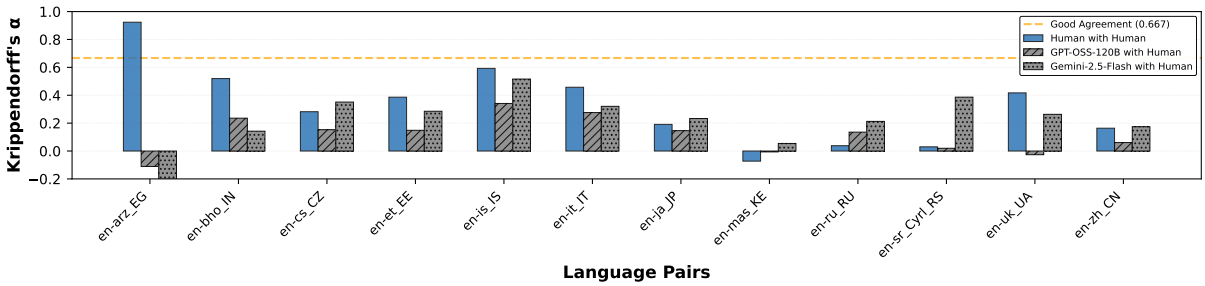
LP	Segs	Sys	Ann	Len
en_US-de_DE	190	13	10	144^{w}
zh_CN-en_US	179	16	8	264^{c}



(a) Dialogue domain.

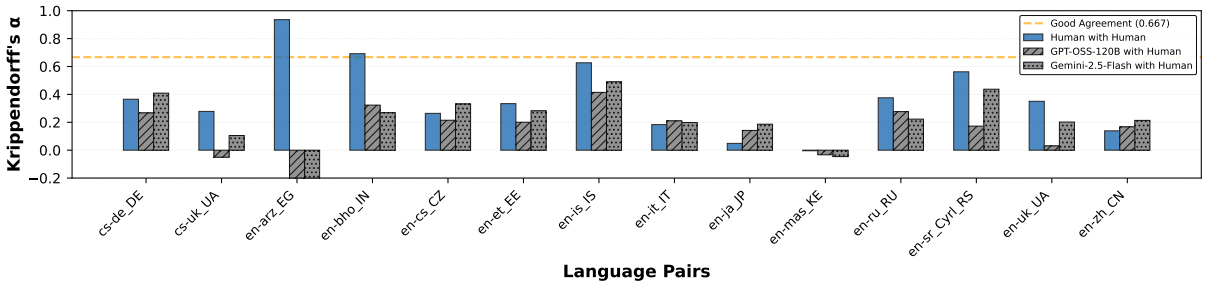


(b) Education domain.

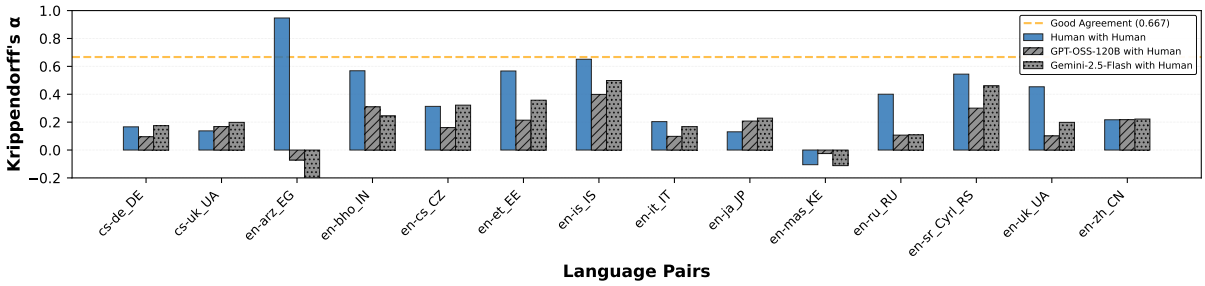


(c) Literary domain.

Figure 15: Inter-annotator agreement for ESA scores on the WMT25 ESA test set across language pairs, for (a) Dialogue, (b) Education, and (c) Literary domains. Bars show Krippendorff's α for each annotator combination.



(a) News domain.



(b) Social domain.

Figure 16: Inter-annotator agreement for ESA scores on the WMT25 ESA test set across language pairs, for (a) News and (b) Social domains. Bars show Krippendorff's α for each annotator combination.