

Post-hoc Correction of Machine Translation Error Span Predictions

Benjamin Pong

University of Washington

Seattle WA, USA

{benpong}@uw.edu

Abstract

This paper presents an unsupervised quality estimation metric for machine translation that combines both neural and LLM models to identify and classify error spans for machine translations. As a submission to the WMT2026 Automated Evaluation Shared Task 1, this system builds on XCOMET-XL by applying offsets to the logits produced by the neural metric, post-hoc, to address the long-tailed probability problem, and uses an LLM as a second-stage to refine these predictions. Results show that XCOMET-XL with logits offset alone shows a boost in recall and character-level F1 scores for error span classifications across multiple language pairs, surpassing state-of-the-art baseline (i.e XCOMET-XL) and LLM-based approaches. This system also comprises a second-stage where a reasoning LLM-judge is employed to audit the error spans predicted by the logits-adjusted XCOMET metric, whose goal is to reduce spurious error classifications and improve overall precision. Results show that using LLM-judge to refine error spans provides no measurable improvements over predictions of the logits-adjusted XCOMET metric.

1 Introduction

This paper presents a quality estimation metric for machine translation that combines both neural and LLM models to identify and classify error spans for machine translations¹. It is well established that classification problems exhibit long-tailed distribution where label distribution is imbalanced. In the context of MT error span detection, state-of-the-art metrics like XCOMET are designed to classify subword tokens based on their error categories. While

there are rigorous attempts to balance the label distribution of error categories in the training data used to train a metric like XCOMET (Guerreiro et al., 2024), majority subwords are still fundamentally classified as no-error, suggesting that the long-tailed distribution problem still exists in the training data. The system builds on XCOMET-XL (Guerreiro et al., 2024) by applying offsets to the logits produced by the neural metric, post-hoc, to address the long-tailed probability problem. To our knowledge, the application of logits adjustments have not been explored in the context of Machine Translation Error Span detection.

This system also comprises a second-stage where a reasoning LLM-judge is employed to audit the error spans predicted by the logits-adjusted XCOMET metric to reduce spurious error classifications. The reason for employing reasoning LLM models for refinement is due to evidence that reasoning models are able to guide the model, through reasoning traces, to arrive at an informed response, for machine translation quality estimation (Maheswaran et al., 2025). Given that error span annotation is a complex task as it requires identifying span indices, evaluation of translation quality, and classification of error types, this task could benefit from reasoning models. However, there is also alternative evidence that LLMs are not entirely good at information extraction tasks (Ma et al., 2023), and may work better when presented with candidates to assess. Motivated by this research, this system explores using state-of-the-art neural metrics and LLMs as refiners of upstream auto-raters.

2 Model Architecture

2.1 XCOMET with logits adjustments

To address the long-tailed probability problem, we adopt the approach by Menon et al. (2021), where the label frequencies are computed across XCOMET’s training data, and the logits produced

¹Code is available at <https://github.com/Benjamin-Pong/wmt2026>

by the neural metric are then re-weighted based on these label frequencies.

We consolidate three MT quality-evaluation datasets that were originally used to train XCOMET-XL (Guerreiro et al., 2024): IndicMT (Singh et al., 2024), WMT (2020–2022), and DEMETR (Karpinska et al., 2022) into a unified corpus. To characterize label imbalance in the consolidated corpus, we computed the distribution of severity labels at the subword level. We tokenize each machine translation prediction in the corpus with the XCOMET-XL encoder, projecting the character-level error labels onto subword tokens to obtain per-token labels drawn from a four-level taxonomy (no-error, minor, major, critical). We aggregated the token labels globally and per language pair, and report each severity class as a fraction of all subword tokens, both overall and per language pair (See Table 1).

Following the logit-adjustment framework of Menon et al. (2021), we offset each class logit by $k \cdot \log(p)$, where p is the class frequency, so that low-frequency error classes receive a relative boost. For simplicity we use the global label frequencies as the prior, with no language-pair-specific weighting, and set $k = 0.5$. In our experiments, larger values collapse the system to predicting no error spans (recall $\rightarrow 0$, and hence F1 $\rightarrow 0$, with precision undefined/1 by convention), so 0.5 is the practical upper bound at which the adjustment still yields useful error detection.

These label frequencies were then applied, post-hoc, to the logits produced by XCOMET on the development and test data. Additionally, XCOMET was also adjusted to produce the per-span entropy at inference time. The purpose of computing entropy is to rank the error spans for further refinement by a reasoning LLM. The reason for refinement is motivated by the results (see Table 2), where the logit adjustments do not show much of a boost in precision, suggesting that there is still the issue of spurious generation of error spans. The goal of the LLM refinement stage is to prune spurious error span predictions to improve precision.

2.2 Reasoning LLMs for refinement

The error spans predicted by logits-adjusted XCOMET-XL were inputs into a reasoning LLM model through an extended prompt. We used DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025) as the LLM-as-judge served with vLLM, decoding greedily (temperature 0) for deterministic,

Source	LP	No-err.	Minor	Major	Crit.
<i>Global</i>		88.38	5.35	5.66	0.61
WMT	en-de	88.82	7.70	3.48	0.00
	zh-en	84.56	5.97	9.47	0.00
	en-ru	91.35	3.61	3.65	1.39
DEMETR	zh-en	94.62	1.85	0.73	2.80
	de-en	94.21	1.95	0.82	3.02
	hi-en	94.51	1.84	0.77	2.88
	ja-en	94.51	1.86	0.73	2.90
	pl-en	94.56	1.87	0.79	2.78
	ru-en	94.39	1.91	0.83	2.87
	cs-en	94.60	1.82	0.78	2.81
	fr-en	94.72	1.79	0.69	2.80
	es-en	94.44	1.87	0.75	2.94
	it-en	94.66	1.80	0.75	2.79
IndicMT	en-gu	97.99	0.67	1.16	0.17
	en-hi	97.32	0.02	1.09	1.57
	en-ml	97.71	0.25	1.65	0.40
	en-mr	97.61	0.10	0.70	1.60
	en-ta	96.04	0.39	1.65	1.92

Table 1: Subword-level severity label distribution (%) in the consolidated corpus, reported globally and per language pair.

reproducible output. We conducted two ablations experiments in the selection of error spans for LLM refinement. First, we experimented with error span prediction through entropy-ranking criteria. Rather than re-judging every span, we rank each system’s XCOMET error spans by their per-span entropy in descending order and pass only the top-k ($k=5$) highest-entropy spans to the judge. The intuition is that entropy serves as a proxy for XCOMET’s uncertainty, so this focuses the LLM’s effort on the spans it is least confident about. Conversely, low-entropy spans are carried through unchanged. The judge’s corrected top-k spans are then re-merged with the untouched spans to form the refined error set. Next, to test if this ranking approach truly makes a difference, we conducted an additional experiment where k randomly selected spans were refined ($k=5$).

Our extended prompt builds on GEMBA (Kocmi and Federmann, 2023), with explicit reasoning steps to refine error spans produced by XCOMET or any other autorater. The prompt supplies the source language, source segment, target language, machine translation, and the candidate error spans, together with the MQM error taxonomy (accuracy, fluency, locale convention, style, terminology, non-translation, other, no-error) and definitions of the major/minor severity levels. The model is instructed to reason step by step over each predicted span and take one of four actions: confirm it, cor-

Language Pair	Baseline			WMT2025 Submissions									XCOMET-XL _{logitsadj}		
	XCOMET-XL			AutoLQA			AIP			GemSpanEval					
	P	R	f1	P	R	f1	P	R	f1	P	R	f1	P	R	f1
CS-DE_DE	24.55	5.15	8.52	17.71	4.02	6.56	11.94	20.37	15.06	28.89	6.02	9.96	6.09	29.23	10.07
CS-UK_UA	25.02	4.00	6.90	20.73	1.93	3.54	13.60	10.16	11.63	35.94	3.58	6.52	3.33	38.21	6.13
EN-AR_EG	11.57	19.53	14.54	11.45	22.95	15.28	2.51	30.41	4.63	19.23	22.18	20.60	40.23	28.49	33.35
EN-BHO_IN	22.87	3.47	6.02	15.55	3.48	5.69	9.38	8.19	8.74	28.40	3.68	6.52	67.48	16.18	26.10
EN-CS_CZ	16.06	6.02	8.76	14.22	4.12	6.39	7.27	15.30	9.85	24.20	6.38	10.10	8.09	37.24	13.30
EN-ET_EE	14.93	16.66	15.75	14.84	11.82	13.16	5.71	20.34	8.92	23.09	12.28	16.04	25.73	35.88	29.97
EN-IS_IS	22.41	27.72	24.78	8.85	19.68	12.21	9.15	35.69	14.57	25.90	19.58	22.30	32.16	35.87	33.91
EN-IT_IT	30.60	4.16	7.33	17.89	2.55	4.47	10.45	13.70	11.86	33.71	5.47	9.41	4.64	44.24	8.39
EN-JA_JP	13.97	3.67	5.81	22.70	2.30	4.18	8.88	11.72	10.10	28.47	3.32	5.94	3.77	38.77	6.86
EN-KO_KR	8.74	14.64	10.95	20.23	7.26	10.69	4.81	25.89	8.12	17.65	10.54	13.20	-	-	-
EN-MAS_KE	10.23	34.84	15.81	15.14	38.95	21.80	27.94	28.65	28.29	35.03	35.67	35.35	88.55	5.18	9.78
EN-RU_RU	16.70	8.59	11.34	13.77	3.84	6.01	8.74	16.77	11.49	28.28	6.49	10.55	8.37	40.42	13.87
EN-SR_CYRL	21.18	21.59	21.38	13.61	15.16	14.35	6.81	27.11	10.88	21.66	15.67	18.18	24.25	37.58	29.48
EN-UK_UA	21.84	2.98	5.25	15.48	1.32	2.43	12.63	6.98	8.99	37.01	2.19	4.13	2.55	46.23	4.83
EN-ZH_CN	22.62	3.80	6.50	13.08	2.92	4.78	7.72	10.43	8.87	30.02	3.37	6.07	3.24	48.45	6.08
JA-ZH_CN	26.89	21.80	24.08	14.83	13.58	14.18	8.47	41.64	14.08	25.35	17.46	20.68	-	-	-
Average	19.39	12.41	12.11	15.63	9.74	9.11	9.75	20.21	11.63	27.68	10.87	13.47	22.75	34.43	16.58

Table 2: Character-level precision, recall, and F1 (%) per language pair across the submissions for WMT2025, and our submission. The average F1, P, R are computed over all except JA-ZH and EN-KO. Bold marks the best F1, P, R per language pair.

rect its severity if misclassified, delete it if it is not a genuine error, or re-segment it into finer sub-spans when only part of the span is erroneous. The output is constrained to a JSON array of span objects with fixed keys (text, start, end, severity), enabling deterministic parsing back into the evaluation pipeline. See Appendix A for the prompt template.

3 Experiments

We developed and evaluated our system on the official test set from WMT2025 Automated Metrics Shared Task (Lavie et al., 2025), and benchmarked our system against the official submissions of WMT2025 Automated Metrics Shared Task Track 2.

We compared four systems, spanning the baseline metric, our proposed adjustment, and two refinement ablations:

- **XCOMET-XL**: the baseline metric, used without modification.
- **XCOMET-XL_{logitsadj}**: our system, applying the logit adjustment to shift the error-decision boundary.
- **XCOMET-XL_{logitsadj} + LLM refine**: an ablation adding LLM-based refinement of randomly selected error spans.
- **XCOMET-XL_{logitsadj} + LLM refine + en-tRank**: an ablation adding LLM-based refine-

ment of entropy-ranked error spans.

Metrics Identical to the evaluation strategy for the WMT2025 Automated Metrics shared task, the evaluation of the automated metrics uses a character-level micro-F1 between predicted error spans by our systems and the gold error spans by human raters. The metric awards full credit when a character’s error is both located and classified correctly, and half credit when the span is identified at the right position but assigned the wrong severity. F1 is computed over multiple overlapping annotations per character, so that co-occurring error spans are scored independently rather than collapsed into one. Concretely, for every hypothesis we tally major and minor errors at each character position separately, for both the gold labels and the predictions.

4 Results and Discussion

4.1 Impact of logits adjustment

Table 2 shows the results for XCOMET-XL with post-hoc logits adjustments. Logits adjustment showed an overall boost recall and F1 across majority of the language pairs, and a modest boost for certain language pairs compared to the XCOMET-XL baseline. For some language pairs like CE-DE and CS-UK, the boost in Recall followed a drop in Precision. XCOMET-XL with logits adjustments also surpassed LLM-powered autoraters like AutoLQA (Hrabal et al., 2025), AIP (Yeom et al., 2025) and

Language Pair	XCOMET-XL _{logitsadj}								
	XCOMET-XL _{logitsadj} + LLM refine						XCOMET-XL _{logitsadj} + LLM refine + entRank		
	P	R	f1	P	R	f1	P	R	f1
CS-DE_DE	6.09	29.23	10.07	6.43	25.22	10.25	6.43	25.24	10.25
CS-UK_UA	3.33	38.21	6.13	3.49	32.02	6.29	3.51	32.36	6.34
EN-AR_EG	40.23	28.49	33.35	41.54	23.86	30.31	41.53	23.81	30.27
EN-BHO_IN	67.48	16.18	26.10	46.94	14.50	22.15	47.57	14.49	22.21
EN-CS_CZ	8.09	37.24	13.30	8.33	31.82	13.20	8.32	31.78	13.19
EN-ET_EE	25.73	35.88	29.97	26.49	30.53	28.37	26.53	30.59	28.42
EN-IS_IS	32.16	35.87	33.91	32.10	29.91	30.97	32.01	29.86	30.90
EN-IT_IT	4.64	44.24	8.39	5.65	44.44	10.02	5.59	43.95	9.91
EN-JA_JP	3.77	38.77	6.86	3.89	32.98	6.95	3.90	33.00	6.97
EN-MAS_KE	88.55	5.18	9.78	83.05	9.02	16.28	84.14	9.26	16.68
EN-RU_RU	8.37	40.42	13.87	8.83	35.25	14.12	8.86	35.41	14.18
EN-SR_CYRL	24.25	37.58	29.48	23.99	30.06	26.68	23.96	30.03	26.65
EN-UK_UA	2.55	46.23	4.83	2.80	42.18	5.25	2.79	42.00	5.23
EN-ZH_CN	3.24	48.45	6.08	3.39	38.10	6.23	3.24	48.45	6.08
Average	22.75	34.43	16.58	21.21	29.99	16.22	21.31	30.73	16.23

Table 3: Character-level precision, recall and F1 (%) per language pair for the logits-adjusted system and two refinement ablations: LLM refinement of randomly selected spans (+ LLM refine) and of entropy-ranked spans (+ LLM refine + entRank).

System	Major			Minor		
	P	R	f1	P	R	f1
XCOMET-XL	15.13	28.21	12.18	12.55	1.31	1.88
XCOMET-XL _{logitsadj}	18.33	36.03	12.95	10.64	13.92	8.23
+ LLM refine	18.55	32.47	12.80	11.35	11.71	8.12
+ LLM refine + entRank	18.61	33.20	12.80	11.52	11.98	8.21

Table 4: Macro-averaged per-severity character-level P/R/F1 (%) over the 14 ESA language pairs, for major and minor errors scored as independent channels.

GemSpanEval (Juraska et al., 2025). This result validates the use of fast and efficient unsupervised calibration approach compared to training a new model for error span classification.

This boost in recall could be attributed to the upweighting of less represented error categories, hence assigning actual error categories to subwords rather than no-error categories. However, as mentioned, there are some tradeoffs in precision for certain locales.

4.2 Impact of LLM-judge refinements

Table 3 shows that LLM refinement yields a small but consistent drop in F1 relative to the logits-adjusted system, holding across both severity levels. The drop reflects a precision–recall trade: refinement slightly raises precision while reducing recall, for a net-negative effect. Selecting spans to refine by entropy rather than at random makes no measurable difference (F1 within 0.01 on average), suggesting that the limitation lies in the refinement step itself rather than in which spans are chosen. A likely contributor is that the base spans already

achieve high recall, leaving refinement more scope to alter correct spans than to recover missed ones

4.3 P, R and F1 by Severity Label

Aforementioned, the logits-adjusted XCOMET-XL showed a notable boost in recall. To investigate where the boost in recall is concentrated, we computed the average Precision, Recall and F1 for each severity label (minor/major). Based on the results, as seen in Table 4, the logits-adjusted XCOMET-XL achieves higher recall and F1 on major errors than on minor errors. The recall gained by the logit adjustment is therefore concentrated on the more severe, higher-priority errors rather than on minor ones. In a similar vein, we also observe a boost in precision due to logits adjustments for major errors albeit a drop in precision for minor errors.

While LLM refinement produced small gains in precision on major errors (18.33 to 18.61), consistent with the step pruning spurious spans, this came at the expense of recall (36.03 to 33.20), and the net effect on F1 was slightly negative (12.95 to 12.80). The same precision-for-recall trade holds for minor

errors, again yielding no net gain in F1. Selecting spans by entropy rather than at random produced near-identical results across both severities, indicating that the ranking provides no measurable benefit over random selection.

5 WMT2026 Submission

We participated in the WMT2026 task using the logits-adjusted XCOMET-XL with LLM refinements on entropy-ranked error spans, on the eight high-priority language pairs² that were designated by the Organizers. As the 2026 edition’s gold annotations for the test set had not been released at the time of submission, the results cannot be reported. All experiments in this paper are conducted on the WMT2025 test set, in accordance with WMT2025’s evaluation script.

6 Conclusion

This system proposes the use of logits offsets to rebalance the label distribution, addressing the long-tailed probability problem for machine translation quality estimation of error spans. Our results show that a simple post-hoc adjustment can result in a boost in F1 and recall, but dampening precision for some locales.

7 Limitations & Future Work

One limitation of this work is that it has yet to address language specific behavior. Deeper statistical investigation needs to be done to find out why the boost in precision and recall from logits adjustments is not uniform across all languages, and why precision benefits little from logits adjustments.

Cost in terms of high token usage is a further consideration: refining every span for every system with a 32B reasoning model is expensive, and because we already rank spans by uncertainty, refinement could be applied adaptively to only the least-confident spans, or delegated to a smaller distilled judge.

A further limitation is that entropy (computed from xcomet) was the sole metric used to rank error spans that the model is presumably least certain about. Future work could experiment with using other metrics like confidence and log-likelihood, or a combination of them, to rank error spans for refinement.

²<https://www2.statmt.org/wmt26/mteval-task.html>

8 Acknowledgments

We thank the reviewers for their detailed and thoughtful feedback.

References

- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. [CUNI and phrase at WMT25 MT evaluation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 934–944, Suzhou, China. Association for Computational Linguistics.
- Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. [MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. Association for Computational Linguistics.
- Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. [Demetr: Diagnosing evaluation metrics for translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Tom Kocmi and Christian Federmann. 2023. [Gembamqm: Detecting translation quality error spans with gpt-4](#). Preprint, arXiv:2310.13988.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, 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.
- Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. [Large language model is not a good few-shot information extractor, but a good reranker for hard samples!](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

- Monishwaran Maheswaran, Marco Carini, Christian Federmann, and Tony Diaz. 2025. [TASER: Translation assessment via systematic evaluation and reasoning](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1004–1010, Suzhou, China. Association for Computational Linguistics.
- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. [Long-Tail Learning via Logit Adjustment](#). In *International Conference on Learning Representations*.
- Anushka Singh, Ananya Sai, Raj Dabre, Ratish Pudupully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. [How good is zero-shot MT evaluation for low resource Indian languages?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.
- Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and Jinyeong Bak. 2025. [Tagged span annotation for detecting translation errors in reasoning LLMs](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 878–886, Suzhou, China. Association for Computational Linguistics.

A Prompt Template for Refinement

Refinement Prompt

You are an annotator for the quality of machine translation. Your task is to reflect on the predicted error spans produced by an autorater given the source and target segments, and re-segment and/or re-classify these spans by severity.

```
source language: {source_language}
source segment: {source_segment}
target language: {target_language}
machine translation: {target_segment}
predicted error spans: {xcomet_error_spans}
```

Guidelines and Error Taxonomy

Error categories: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, time format), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.

Each error is either major or minor. Major errors inhibit comprehension or disrupt flow. Minor errors are technically errors but do not disrupt flow or hinder comprehension. Formatting errors alone do not hinder comprehension.

Your Task

Evaluate only the predicted error spans, given the machine translation in {target_language} and the source segment in {source_language}. Think step by step:

1. Each entry has 'text' from the machine translation, predicted as a mistranslation. Decide whether it is a true error, referring to the source.
2. If it is a true error but its severity is wrong, reclassify the severity.
3. If the span is not an error at all, remove its entry from the array.
4. If only part of the span is an error, re-segment it into subspans and emit those instead, each with keys: {'text': str, 'start': int, 'end': int, 'severity': str}.

Output a JSON array of objects, each with exactly the keys "text", "start", "end", "severity". 'start'/'end' are character indices into the target translation, 'text' is the substring they bound, and "severity" is "major" or "minor". If nothing is an error, return []. Output ONLY the JSON array, with no other text.

Figure 1: Reference-free refinement prompt that instructs an LLM to audit the error spans predicted by an upstream autorater.