Can QE-informed (Re)Translation lead to Error Correction?

Govardhan Padmanabhan

Institute for People-Centred AI
University of Surrey, United Kingdom
gp00816@surrey.ac.uk

Abstract

The paper presents two approaches submitted to the WMT 2025 Automated Translation Quality Evaluation Systems Task 3 - Quality Estimation (QE)-informed Segment-level Error Correction. While jointly training QE systems with Automatic Post-Editing (APE) has shown improved performance for both tasks, APE systems are still known to *overcorrect* the output of Machine Translation (MT), leading to a degradation in performance. We investigate a simple training-free approach - QE-informed Retranslation, and compare it with another within the same training-free paradigm. Our winning approach selects the highest-quality translation from multiple candidates generated by different LLMs. The second approach, more akin to APE, instructs an LLM to replace error substrings as specified in the provided QE explanation(s). A conditional heuristic was employed to minimise the number of edits, with the aim of maximising the Gain-to-Edit ratio. The two proposed approaches achieved a ΔCOMET score of 0.0201 and -0.0108, respectively, leading the first approach to achieve the winning position on the subtask leaderboard.

1 Introduction

Large Language Models (LLMs) have advanced the field of Machine Translation (MT), given their support for longer input context length and ability to generate text in a natural tone. However, translation quality is still limited for languages other than English and for domain-specific translations (Fernandes et al., 2025). Evaluating MT output for quality is critical to understand the reliability and suitability of translation systems, and more importantly, to be able to perform accurate corrections. The WMT24 Metrics Shared Task found that neural-based learned metrics like COMET or xCOMET were superior when evaluating LLM-generated translations, compared to the traditional statistical metrics like BLEU, or chrF (Freitag et al., 2024).

As LLMs are typically trained on a large general dataset, performance in domain-specific MT can often fall short as they may not properly render key terminologies or stylistic conventions. As such, Automatic Post-Editing (APE) is vital to fixing MT errors. However, they are prone to overcorrecting. One such example of mitigating over-correcting, proposed by Deoghare et al. (2025), is to utilize word-level Quality Estimation (QE) to limit edits only on the specified error segments. Despite recent efforts to reduce over-correction (Deoghare et al., 2023, 2024), APE models still fall short of the required semantically coherent output. Therefore, we ask the titular question - "Can QE-informed re-translation help overcome MT errors?", and discuss two approaches as a comparative evaluation.

This paper describes two participation systems, both utilizing pre-trained and open-source LLMs, for the WMT 2025 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. The primary approach leverages multiple LLMs for MT and selects the best output using QE. In the secondary approach, an LLM is prompted to replace error-segments in the provided MT. These error segments are identified by the explainable QE provided within the dataset.

2 Related Work

Quality Estimation (QE) QE is an automated evaluation framework that predicts a score indicating whether the translation is good or not (Yvon, 2019). COMET (Cross-Lingual Optimized Metric for Evaluation of Translation) is an automatic metric to evaluate the quality of machine translation using deep learning models (Rei et al.,

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 984–993
November 8–9, 2025.

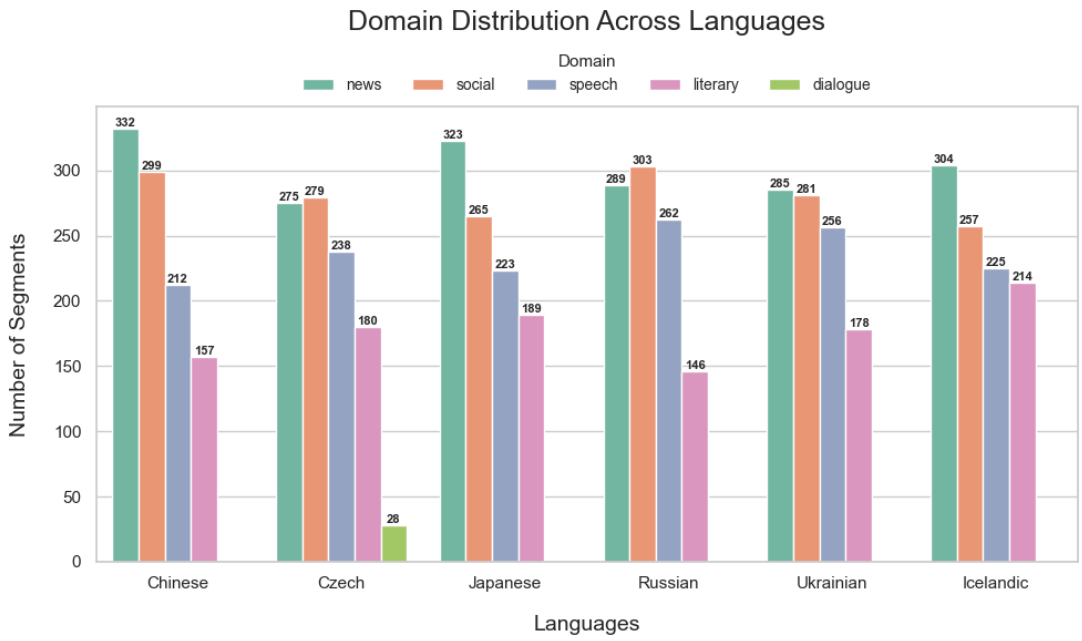

Figure 1: Domain distribution within each language

2020). COMETKiwi is a hybrid machine translation quality estimation (MTQE) model that combines COMET and OpenKiwi (Rei et al., 2022). It achieved top performance in the WMT 2022 shared task and has since been widely adopted as a state-of-the-art benchmark in MTQE.

Automatic Post-Editing (APE) APE is an automated system that corrects MT output, without human involvement (do Carmo et al., 2021). Chatterjee et al. (2018) describe combining QE and APE in three ways: using QE as an APE activator when the MT output is poor, as guidance to help the APE decoder decide which tokens to change, and as a selector to choose between the raw MT and post-edited output.

WMT24 QE-APE The previous year's WMT competition focused on sentence-level quality estimation and error span predictions (Zerva et al., 2024). QE was further incorporated into APE. The dataset for the QE-APE primarily consisted of English–Hindi (En-Hi) and English–Tamil (En-Ta) pairs. The source (SRC) English sentence, the target (TGT) translation provided by an unspecified neural machine translator, and a human post-edit (PE) version of the translation made by native speakers were included to support both quality estimation and automatic post-editing tasks.

The HW-TSC team (Yu et al., 2024) utilized `Llama3-8B-Instruct` for En-Hi pairs. The model first underwent continual pretraining using low-rank adaptation (LoRA) on SRC and TGT data, and was further supervised fine-tuned on the PE data with a custom prompt. For En-Ta pairs, they trained a custom transformer then performed APE fine-tuning. This system achieved 0.851 and 0.918 COMET scores for En-Hi and En-Ta translation tasks, respectively.

The IT-Unbabel team utilized `xTower` for generating corrected translations, along with a quality estimation model to decide whether to use the original translation or the `xTower` output. This approach achieved 0.8646 COMET score for En-Hi pairs, and 0.9163 for En-Ta pairs.

IT-Unbabel's solution of utilizing an LLM along with QE as a selector served as the inspiration for the primary approach

3 Methodology

3.1 Dataset Analysis

The complete test data provided with the task was exclusively used. This dataset consists of 6,000 machine translations from English to Chinese, Czech, Japanese, Icelandic, Russian, and Ukrainian, with 1,000 instances for each language pair. The texts cover a range of domains, specifically news, social, speech, literary, and dialogue. As shown in Figure 1, the domains are not evenly distributed. The *news* domain is the most prominent overall with 1,808 entries, while the *dialogue*

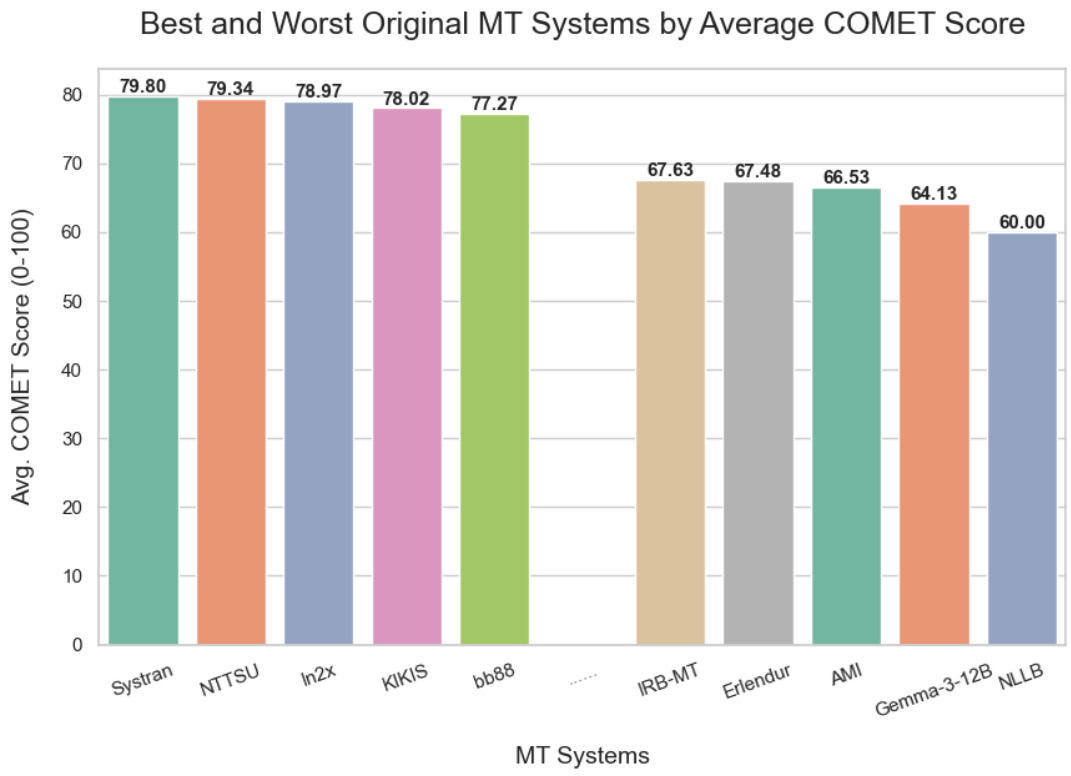

Figure 2: System performance in hypothesis_segment

domain has the lowest representation with 28 segments, all appearing only in the English-to-Czech group.

The original translations under the "hypothesis_segment" column were generated by 38 different translation systems, including LLMs like `GPT4.1`, `Claude-4`, `DeepSeek-v3`, and more. Figure 2 shows the five best and worst systems according to their average translation quality scores.

3.2 Approaches

3.2.1 Primary Approach - "Best MT Wins"

In this approach, multiple LLMs were used to translate the English texts from scratch, without additional information or context. The resulting candidate translations were then evaluated with the *wmt22-cometkiwi-da* model, which provided QE scores based on the source English text and translated system outputs. The translation with the highest QE score was selected as the final output. By re-framing the APE task as re-translation, QE serves as a selector in place of traditional decoders. Similar approaches have also been explored, where QE is used in multi-hypothesis selection (Yu et al., 2024; Laki and Yang, 2018; Lu and Zhang, 2019), further supporting the view of QE as a decision mechanism in MT improvement. This approach therefore functions as a systematic probe of QE-based re-ranking, illustrating both its potential to establish an empirical upper bound on metric-based performance and its limitations in terms of computational cost and efficiency.

`Aya-Expanse-8B`, `GPT-SW3-6.7B`, `Tri-7B`, `GLM4-9B`, `Phi4-mini-instruct`, and `TowerPlus-9B` models were used. These models were selected because of their reported performance, robustness, popularity, and recency. Models like `GPT-SW3-6.7B`, `Tri-7B`, and `GLM4-9B` were selected owing to their specialized training in certain languages, specifically Icelandic, Japanese, and Chinese respectively.

The prompts for `Tri-7B`, `Phi4-mini-instruct`, and `TowerPlus-9B` are reminiscent of their translation prompts available in their Huggingface model cards, being variations of:

> Translate from English to *{Language}*

System prompts for the other models were more involved, as using the same prompt resulted in some hallucinations, chain of thought, or worse translation quality during limited internal testing.

For `Aya-Expanse-8B`:

> You are a helpful bilingual assistant that correctly translates the user's input text from English to *{Language}*.
> When translating, you must use the same tone and intent of the English text. You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.

For `GPT-SW3-6.7B`:

> <|endoftext|><s>
> System:
> You are a bilingual assistant that objectively translates the user's input text from English to Icelandic.
> You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.
> <s>
> User:
> *{original English text}*
> <s>
> bot:

For `GLM4-9B`:

> You are a bilingual assistant that correctly translates the user's input text from English to Chinese.
> When translating, you must use the same tone and intent of the English text. You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.
> Output only the translation!

Except `TowerPlus-9B`, the remaining models do not support all required languages. Hence, they translated only their supported language(s), and the rest were omitted. After all models were successfully executed, the QE score via COMETKiwi between all translations, including the original systems', was used to determine which MT to use.

3.2.2 Secondary Approach - "Fill in the Blanks"

The provided test data includes *error spans* for the translations. Using fine-grained QE signals to guide targeted corrections, this approach investigates whether restricting edits to QE-highlighted segments could yield improvements with fewer changes, thereby increasing both efficiency and interpretability compared to full re-translation.

Using these error spans, the corresponding substring(s) in translation is replaced with a "**__BLANK__**" token, emulating a multilingual masked language modeling task. The text domain is also used to provide additional context. An example is included in the prompt to guide the model's behavior, serving as a one-shot example. Different examples were used in the prompt for different languages.

This approach utilizes `TowerPlus-9B`, an open source LLM with 9 billion parameters based on Gemma2. This model was selected because of its training for translation-related tasks along with instruction tuning, a context window of 8192 tokens, and has support of the languages of this task: English, Chinese, Czech, Japanese, Icelandic, Russian, Ukrainian. It has also been shown to outperform larger parameter models like `Gemma2-27B` and `Llama3.3-70B` on translation performance (Rei et al., 2025). The 9B variant was specifically selected due to resource and time constraints.

Provided below is the system prompt template for Russian language translations:

> You are a helpful assistant that corrects a Russian translation by filling in the blanks. Use the English sentence for context. Complete the task while maintaining the tone of a *{domain}*.
> Important - Do not use any of the specified wrong words. Replace each __BLANK__ token with an appropriate word or phrase that matches the original meaning, tone, and context of the English sentence.
>
> Example (social domain):

Russian with __BLANK__: Многие молодые люди сегодня предпочитают __BLANK__ в кофейнях, а не дома.
English: Many young people today prefer to hang out in coffee shops rather than at home.
Wrong translation: Многие молодые люди сегодня предпочитают учиться в кофейнях, а не дома.
Wrong words: [' учиться']
Corrected Russian sentence: Многие молодые люди сегодня предпочитают проводить время в кофейнях, а не дома.

Russian with __BLANK__: *{mt with __BLANK__}*
English: *{source text}*
Wrong translation: *{mt text}*
Wrong words: *{list of substrings removed}*
Corrected Russian sentence:

To increase translation quality while keeping changes minimal (gain-to-edit ratio), a conditional masking heuristic based on error severity and overall QE score was employed. The pseudocode for this method is provided in Algorithm 1.

Algorithm 1 Conditional masking based on severity and score

```
Original QE Score: Float x
if x >= 0.90 then
    Proceed without masking
else if x > 0.50 then
    if only minor severity error spans then
        Mask minor error spans
    else
        Mask all non-minor error spans
    end if
else
    Mask all error spans
end if
```

4 Results and Discussion

Tables 1 and 3 present language-wise results for the primary and secondary approaches, respectively. The WMT'25 shared task tracks two main metrics: ΔCOMET and Gain-to-Edit Ratio (referred to as "G2E Ratio"), with the overall ΔCOMET score serving as the primary selection criterion. Here, ΔCOMET is computed as the difference between the COMET scores of the proposed approach and the baseline translation. The G2E Ratio is calculated as ΔCOMET divided by the total edit rate. BLEU and chrF++ scores are also included in the tables as supplementary information, but they are not the focus of the analysis. The best-performing ΔCOMET and G2E Ratio scores across the two tables are highlighted in bold.

4.1 Primary Approach - "Best MT Wins"

Language	ΔCOMET	G2E Ratio	BLEU	chrF++
Icelandic	$3.65e-2$	$1.27e-3$	69.24	78.37
Russian	$2.01e-2$	$7.10e-4$	68.56	79.15
Czech	$1.89e-2$	$8.15e-4$	73.85	82.29
Chinese	$1.84e-2$	$1.64e-4$	39.35	58.41
Ukrainian	$1.63e-2$	$6.36e-4$	71.36	80.85
Japanese	$1.03e-2$	$1.41e-4$	54.33	65.44
Average	$\mathbf{2.01e-2}$	$\mathbf{6.22e-4}$	62.78	74.08

Table 1: Language-wise Results for Primary Approach

Table 1 shows that an ensemble of diverse models helps significantly improve the translations, especially in low-resource languages like Icelandic with roughly 3% improvement.

As this approach selects the translation with the best QE score, not all systems or model responses contributed equally to the final output. As shown in Table 2, `TowerPlus-9B` and the original translations have contributed the most, while `Tri-7B` did not contribute at all. Despite `GPT-SW3-6.7B`, `Tri-7B`, and `GLM4-9B` trained primarily for use in Nordic and Altaic languages and Chinese, other models provided better translations. Figure 3 shows how different models contributed to the final response in each language.

System	Contribution
Tower+ 9B	2836
Original	2829
GLM4 9B	149
Phi4 Mini Instruct	106
Aya Expanse 8B	76
GPT-SW3 6.7B v2 Instruct	9
Tri 7B	0

Table 2: System-wise Contribution to the Output

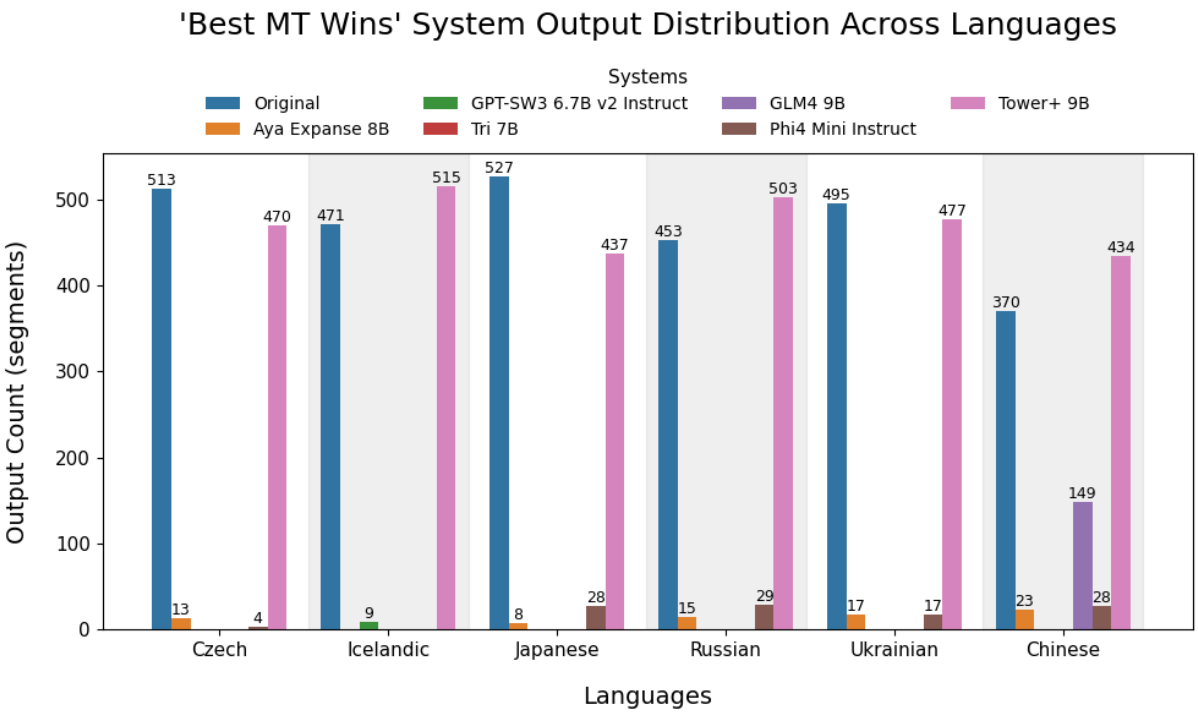

Figure 3: Language-wise count of model responses used in primary approach's final output

The example in Appendix A is an instance where the direct translation by `TowerPlus-9B` was the best-performing. Interestingly, the corresponding output from the secondary approach is of lower quality. This demonstrates the powerful translation capabilities of the model and the need for better prompt engineering in the secondary approach.

4.2 Secondary Approach - "Fill in the Blanks"

Language	ΔCOMET	G2E Ratio	BLEU	chrF++
Czech	$-7.24e-3$	$-5.80e-7$	92.37	95.14
Russian	$-8.03e-3$	$-6.00e-7$	92.56	95.31
Icelandic	$-1.00e-2$	$-8.20e-7$	74.27	82.18
Chinese	$-1.30e-2$	$-4.83e-5$	29.52	83.59
Japanese	$-1.34e-2$	$-4.44e-5$	25.77	82.59
Ukrainian	$-1.35e-2$	$-1.26e-6$	91.77	94.43
Average	$-1.08e-2$	$-1.59e-5$	67.71	88.87

Table 3: Language-wise Results for Secondary Approach

Table 3 shows that this approach does not provide much benefit to the overall translation task, with roughly -0.7% to -1.4% quality degradations.

Appendix C is an example where this approach yields a better translation than the original. The original Czech translation by `DeepSeek-v3` has 3 major errors and 1 minor error. This improvement by the model could be owed to the fact that, on top of its multilingual capabilities, `TowerPlus-9B` was fine-tuned on instruction data from different models, including `DeepSeek-v3`.

There are possible factors why this approach did not perform well: use of bigger and closed-source models in the original MT, use of varied systems and models, inclusion of low-resource languages, correcting only critical + major error spans and ignoring minor error spans in some instances, and more. In the Appendix B example, the original system used Massively Multilingual Neural Machine Translation (MMMT). Although more capable models have been released since then, including `TowerPlus-9B`, output artifacts ("__HEARTBREAK__") and leakage ("Corrected words: [...") affect the translation's quality. This shows that further prompt tuning and output post-processing is needed.

5 Conclusion and Future Work

This paper described two translation approaches submitted to the WMT 2025 QE-informed Segment-level Error Correction task. The first approach used QE as a selector among multiple

LLM translation outputs, resulting in an overall ΔCOMET of 0.0201. The second approach utilized `TowerPlus-9B` exclusively to replace erroneous words in the MT by masking substrings highlighted in the error spans with a blank token, resulting in ΔCOMET of -0.0108. Custom prompts were designed to instruct the model in correcting the translation, similar to a fill-in-the-blank task.

While the first approach showed positive improvements by using QE as a selector, it ultimately depends on the model and system selection. Further exploration of system combinations could yield better performance. The second approach, which performed worse, corrected translations by filling in error segments in the MT. As future work, a two-model system could be explored: a smaller LM to suggest words or phrases for masked tokens via masked language modeling, and a larger LLM to select the most suitable ones to produce a higher-quality translation.

Limitations

Due to limited time and compute resources, the overall experimental design favored n-shot prompting with LLMs for their ease of use and availability of pre-trained weights. Additionally, the model selection was guided by convenience and practical factors such as parameter count, recency, and language compatibility or specificity. These limitations also limited the scope for testing and prompt optimization.

Though "Best MT Wins" reframes APE as retranslation through QE-based selection, this approach is impractical as it requires generating full hypotheses from large models. While the direct translation capabilities of `TowerPlus-9B` even slightly surpassed the original translation system, the other 5 LLMs used were less effective in comparison, resulting in inefficient use of and wasted time and computational resources.

For the "Fill in the Blanks" approach, a lack of proper prompt tuning and post-processing degraded the output quality, indicating that prompt engineering and output handling are important when using LLMs for specific tasks.

More broadly, both approaches rely exclusively on automatic QE metrics such as COMET, which, while effective for shared task evaluation, are primarily trained on English-centric data. The absence of human evaluation limits the ability to validate whether metric-based gains reflect true improvements in translation quality, especially for non-English language pairs.

6 Acknowledgements

My sincerest gratitude to Dr Diptesh Kanojia, Archchana Sindhujan and Sourabh Deoghare from the shared task team for their valuable feedback and guidance, encouragement, paper review, and assistance with LaTeX formatting.

References

Rajen Chatterjee, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 26–38, Boston, MA. Association for Machine Translation in the Americas.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2024. Together we can: Multilingual automatic post-editing for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10800–10812, Miami, Florida, USA. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2025. Giving the old a fresh spin: Quality estimation-assisted constrained decoding for automatic post-editing.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya. 2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698, Singapore. Association for Computational Linguistics.

F. do Carmo, Dimitar Shterionov, Joss Moorkens, Andy Way, Federico Gaspari, and Joachim Wagner. 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35:101–143.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André F. T. Martins, and Graham Neubig. 2025. Do llms understand your translations? evaluating paragraph-level mt with question answering. *Preprint*, arXiv:2504.07583.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami,

Florida, USA. Association for Computational Linguistics.

László János Laki and Zijian Győző Yang. 2018. Combining machine translation systems with quality estimation. In *Computational Linguistics and Intelligent Text Processing*, pages 435–444, Cham. Springer International Publishing.

Jinliang Lu and Jiajun Zhang. 2019. Select the best translation from different systems without reference. In *Natural Language Processing and Chinese Computing*, pages 355–366, Cham. Springer International Publishing.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jiawei Yu, Xiaofeng Zhao, Min Zhang, Zhao Yanqing, Yuang Li, Su Chang, Xiaosong Qiao, Ma Miaomiao, and Hao Yang. 2024. HW-TSC's participation in the WMT 2024 QEAPE task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 541–546, Miami, Florida, USA. Association for Computational Linguistics.

François Yvon. 2019. Quality estimation for machine translation. *Computational Linguistics*, 45(2):391–394.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Example of Primary Approach

Language: Chinese

Original System: Wenyiil

Original English Text:
A brief note on Play Station controllers to round out the 1st party lineup: I never liked them.\nDual Shock 1-3 felt exactly the same to me with exactly the same flaws: Oversized thumbsticks placed too close together, mushy face/shoulder/trigger buttons that seemed to wear out very quickly, and a D-pad that just felt wrong. Dual Shock 4 fixed a lot of problems, but introduced a bunch of new ones. I got a Dual Sense because some friends had a hard time with asymmetrical thumbsticks, and for some reason local stores were selling them for $20 less than DS4 controllers.\nI don't really use this controller: Face buttons are spaced too far apart, my audio in/output keeps defaulting to the controller when it's connected, the bumpers are weirdly tall, and I keep accidentally pressing the track pad when trying to hit menu/share.\nThey are fine guest controllers that have come a long way in terms of design, but for me they're a novelty at best.

Original Chinese MT:
于PlayStation手柄，里提一下，算是第一方手柄容的一个充：我从来都不喜它。\n在我看来，DualShock 1到3代手感完全一，缺点也如出一：杆帽太大且位置太近，面板按、肩和扳机手感很肉，而且似乎特容易磨，十字的感也是不。DualShock 4修正了多，但也引出了一堆新毛病。我了一个DualSense手柄，是因有些朋友用不非称杆，并且出于某原因，本地商店里它的售价比DS4手柄要便宜20美元。\n我其不怎用个手柄：面板按距太，次接手柄的音入/出都会自切去，肩的高度很奇怪，而且当我想按菜/分享会触到触摸板。\n它作客人用的手柄不，在上也确有了足的，但我而言，充其量算是个新奇玩意儿。

New MT:

于 PlayStation 手柄，我想充一点：我从来不喜它。\nDual Shock 1-3 我来感完全一，而且有同的缺点：拇指杆太大，距离太近；部/肩部/扳机按的，似乎很快就会磨；D 方向感不。Dual Shock 4 解决了好多，但也引入了好多新。我了一个 Dual Sense，因有些朋友不称的拇指杆感到不适，而且不知什原因，当地商店的价格比 DS4 手柄便宜 20 美元。\n我很少用个手柄：部按之的距离太，接我的音入/出是默置手柄；冲器很奇怪地高；当我点菜/分享，我是不小心按下了触控板。\n它是不的客用手柄，在上已有了很大的改，但我来，它充其量只是新奇玩意儿。

Original COMET Score: 0.7751

New COMET Score: 0.7797

Approach 1's Edited COMET Score: 0.7723

B Negative Example for Secondary Approach

Language: Japanese

Original System: MMMT

Original English Text:
Kyle nodded, the corners of his mouth twitching as he fought back a smile. "Is that right? Sorry to let you down. Guess it's a good thing I hung up the skates when I did, spare a few more fantasy teams the heartbreak."\nDeclan opened his mouth, then quickly shut it again, heat creeping up the back of his neck. "I mean…your Cup run with the Kings was insane, though. That game seven OT goal? Unreal." He scratched the back of his neck. "So… uh, what are you doing at St. Cassian's? Are you coaching here now? I didn't see you on the website."

Original Japanese MT:
カイルは頷き、微笑みを抑えながら口角をそらせた。「そうですか。がっかりさせちゃってごめんなさい。急いでスケートを引退しておいてよかったと思うのは、他のファンタジーチームがまた心を痛めることを避けるためでしょう。」\nデクランは口を開けてすぐに閉じ、首の後ろに熱さがこみ上げてきた。"いや…キングスとのカップ戦は、ただただすごいね。あの第7戦の延長戦のゴールは…信じられないよ。"彼は首の後ろを掻いた。「それで…あんた、セント・カシアンズでは何をしてるんだ？今はここでコーチをやっているのか？ウェブサイトにはあなたがいる姿が見当たらなかったけど。」

Modified Japanese MT:
カイルは頷き、微笑みを抑えながら口角をそらせた。「そうですか。がっかりさせちゃってごめんなさい。急いでスケートを引退しておいてよかったと思うのは、他のファンタジーチームがまた__HEARTBREAK__を痛めることを避けるためでしょう。」\nデクランは口を開けてすぐに閉じ、首の後ろに熱さがこみ上げてきた。"いや…キングスとのカップ戦は、ただただすごいね。あの第7戦の延長戦のゴールは…信じられないよ。"彼は首の後ろを掻いた。「それで…あんた、セント・カシアンズでは何をしてるんだ？今はここでコーチをやっているのか？ウェブサイトにはあなたがいる姿が見当たらなかったけど。」
Corrected words: ['HEARTBREAK', 'カップ戦']

Original COMET Score: 0.8183

New COMET Score: 0.7314

Error Spans:
$\{'start\_i' : 87, 'end\_i' : 88, 'severity' : 'major'\}$,
$\{'start\_i' : 150, 'end\_i' : 154, 'severity' : 'major'\}$

C Positive Example for Secondary Approach

Language: Czech

Original System: DeepSeek-v3

Original English Text: The duchess and her husband, Prince Harry, have to do something and good luck to them, but it's hard not to study the rollout of As Ever for signs of the widening gap between Meghan's self-image and how the rest of the world sees her. ""As ever,""writes the duchess on Instagram, "means "as it's always been"or some even say "in the same way as always,""the "some"in this sentence apparently referring to the dictionary definition of a two-word phrase that no one has ever had trouble understanding.

Original Czech MT: Vévodkyně a její manžel, princ Harry, musí něco dělat, a přejme jim hodně štěstí, ale nelze si nevšimnout, že uvedení značky As Ever odhaluje rostoucí propast mezi Meghaniným sebepojetím a tím, jak ji vnímá zbytek světa. „„As ever‘,“ píše vévodkyně na Instagramu, „znamená ‚jak to vždy bylo‘, nebo někteří dokonce říkají ‚stejným způsobem jako vždy‘.“ Tito „někteří“ v této větě zřejmě odkazují na slovníkovou definici dvouslovného výrazu, kterému nikdo nikdy nerozuměl.

Modified Czech MT: Vévodkyně a její manžel, princ Harry, musí něco udělat, a přejme jim hodně štěstí, ale nelze si nevšimnout, že uvedení značky As Ever odhaluje rostoucí propast mezi Meghaniným sebepojetím a tím, jak ji vnímá zbytek světa. „„As ever‘,“ píše vévodkyně na Instagramu, „znamená ‚jak to vždy bylo‘, nebo někteří dokonce říkají ‚stejným způsobem jako vždy‘.“ Tito „někteří“ v této větě zřejmě odkazují na slovníkovou definici dvouslovného výrazu, kterému nikdo nikdy neměl problém porozumět.

Original COMET Score: 0.8215

New COMET Score: 0.8317

Error Spans:
$\{'start\_i' : 42,' end\_i' : 53,' severity' : 'major'\}$,
$\{'start\_i' : 174,' end\_i' : 180,' severity' : 'minor'\}$,
$\{'start\_i' : 447,' end\_i' : 467,' severity' : 'major'\}$,
$\{'start\_i' : 468,' end\_i' : 469,' severity' : 'major'\}$