Principled Domain Adaptation for Localisation Quality Estimation in the Absence of Direct Assessments

Anonymous ACL submission

Abstract

Learned quality estimation metrics such as COMETKiwi are widespread and work well for general machine translation evaluation. However, they are known to struggle on unseen domains, limiting their performance in a commercial localisation context. We also show that they are insensitive to some important factors for commercial machine translation, such as accurately translating numbers, or even preserving the correct number of spaces and punctuation in a translation. Further, a key capability for optimisation of machine translation is the ability of quality estimation metrics to accurately rank different translations of a single segment, which is shown to suffer especially from the domain transfer. In the absence of large-scale direct assessment data, we propose principled fine-tuning approaches to bridge the domain gap with even relatively small amounts of post-editing data. Using two different fine-tuning approaches and a simple tokeniser intervention, we create metric variants which prove markedly better at distinguishing preferred post-edits from rejected initial translations in a localisation context.

1 Introduction

Learned neural machine translation metrics are known to struggle in unseen domains (Zouhar et al., 2024). Metrics, and quality estimation models such as COMETKiwi, have been trained primarily on news and wiki text. In real-life localisation workflows, the domains vary from spec sheets to UI copy to legal text. Additionally, data noise is introduced at different points in the pipeline from poor pre-processing to subjective stylistic edits.

As Figure 1 illustrates, quality estimation (QE) is particularly vulnerable to this domain shift. However, QE is also potentially very important for localisation. Due to the nature of post-editing workflows, available references are frequently similar to the hypotheses, stabilising reference-based metrics and showing the desired terminology of the project—while also biasing evaluation towards the originally used machine translation model.

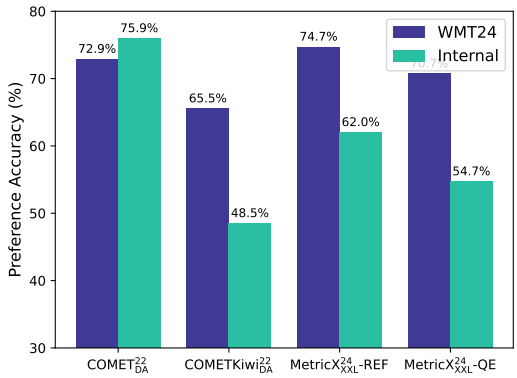

Figure 1: Preference Accuracy of popular metrics and QE models on WMT24 versus our internal test set. Reference-based metrics, particularly COMET$^{22}_{DA}$, cope better with the domain shift than QE models. MetricX$^{24}_{XXL}$ is stronger overall, but its reference-based version seems to orient less strongly towards the reference than COMET$^{22}_{DA}$ does. COMETKiwi$^{22}_{DA}$ is worse than random guessing on the out-of-domain data.

Further, quality estimation is necessary for fine-tuning workflows such as direct quality optimisation (DQO; Uhlig et al., 2025). In this context, an especially important capability is *segment-level ranking* of different translations of the same source, since the quality estimator must reliably pick out preferred translations.

In addition to the domain shift problem, we make several observations which motivate us to adapt learned metrics and particularly QE models:

- For long numbers, incorrect translations are not consistently distinguished and caught.
- Translators and reviewers react strongly to specific issues that the models either cannot see at all—such as the incorrect use of non-

	before	after
Can you confirm that the input 7777777 has a number with seven digits?		
— Können Sie bestätigen, dass die Eingabe 7777777 eine Zahl mit sieben Ziffern hat?	85.91	81.00
— Können Sie bestätigen, dass die Eingabe 777777 eine Zahl mit sieben Ziffern hat?	85.57	76.47
Since they caused the problem, maybe they can do something to _____ it too.		
— Nachdem sie das Problem verursacht haben, können sie vielleicht auch etwas tun, um es zu _____.	78.89	73.42
— Nachdem sie das Problem verursacht haben, können sie vielleicht auch etwas tun, um es zu lösen.	84.20	70.98
These cargo pants are the easiest to dress when you are attending an outing!		
— Ce pantalon cargo est le plus facile à habiller lorsque vous participez à une sortie[NBSP]!	84.83	66.09
— Ce pantalon cargo est le plus facile à habiller lorsque vous participez à une sortie !	84.83	58.20

Table 1: Exemplars of the mentioned issues. The scores shown are from COMETKiwi$^{22}_{\text{DA}}$ ("before") and our domain-adapted model with an adapted tokeniser and fine-tuned with continuous scores on localisation-specific data ("after"). The base model is rather insensitive to errors in long numbers, may prioritise a grammatical sentence over correctly translating a cloze, and cannot see non-breaking spaces.

breaking spaces—, or potentially even reward, such as hallucinating a word to fill a cloze.

Although newer, high-quality QE models have been proposed since its release, COMETKiwi$^{22}_{\text{DA}}$ represents a very good tradeoff between strong performance and a lightweight model. It is both cheap to run and efficient to fine-tune for quick iteration. It has also been found to outperform LLM-based metrics at segment-level ranking (Moghe et al., 2025). Therefore, we use COMETKiwi$^{22}_{\text{DA}}$ as the main model to validate our approach in this work, although our proposed approaches are likely to be more broadly applicable.

Our key contributions are:

1. We formulate a lightweight, data-efficient approach to domain adaptation for QE models, using post-edits and simple heuristics as our training signal. Tokeniser adaptations allow the model to learn specific relevant patterns.

2. We show how to fine-tune a QE model using contrastive learning, yielding much stronger segment-level ranking performance.

3. We demonstrate improvements on localisation-specific issues using our Localisation CheckList challenge set, and on a Human-Evaluated Localisation Dataset (HELD). We aim to release both datasets once this article is de-anonymised.

2 Related Work

2.1 COMETKiwi models

The COMET framework was initially proposed in Rei et al. (2020). In short, COMET models are multilingual encoders which have been finetuned to predict normalised human direct assessments of machine translations. The original COMET models required a reference and only predicted a single sentence-level score, but later iterations expanded to quality estimation (QE) with COMETKiwi (Rei et al., 2022), as well as error span prediction with xCOMET models (Guerreiro et al., 2024). Due to this development history, there are various COMET-style models available and we specify the model names whenever needed.

2.2 Domain Adaptation for Metrics

Similarly to any fine-tuned language model, the COMET models can struggle when applied to unseen domains, as Zouhar et al. (2024) show for reference-based neural models. Sharami et al. (2023) also discuss this phenomenon for quality estimation and use a domain-adaptation approach borrowed from machine translation for their QE model. Thus, they fine-tune a model pre-trained on out-of-domain data with progressively more specific target-domain data. They produce target-domain augmented data by inferencing an MT model and scoring that hypothesis against the existing reference translation via TER (Snover et al., 2006), improving the correlation of their QE model with HTER. The present work instead evaluates both against preferences derived from post-edits and ESA scores from professional annotators. Our approach is also more straightforward, requiring just one fine-tuning step with a mix of target-domain and general data.

Another related approach is "poor man's quality estimation" (Zouhar et al., 2023). They suggest pre-training quality estimation with large-scale *metric estimation* (i.e., predicting the scores of a reference-based metric) before fine-tuning with

small amounts of human labels. In the present work, we instead use metric estimation as a continuous training signal for domain adaptation.

MS-COMET (Kocmi et al., 2022) is a retraining of COMET and COMET-QE with much larger training corpora of 2 M and 3.5 M human-annotated segments, respectively. They cover over 110 languages and 15 domains, using only annotations from professional translators. As part of quality control, they removed campaigns where any of the machine translation systems scores significantly better than the human reference—a choice that may be harder to justify several years later. In contrast to the original COMET models, they use raw direct assessment scores, arguing that this gives more meaning to the score ranges, by assigning languages with generally lower-quality translations some scores in the 0-50 range. However, we note there is a risk of translations into those languages being consistently scored poorly even once they improve, since the model may learn that a particular language usually scores lower. Our approach requires much less data, and we keep z-score normalisation intact.

2.3 Contrastive Training

Contrastive training has been proven effective for representation learning in the past, such as in InfoXLM (Chi et al., 2021) or mSimCSE (Wang et al., 2022). Preference tuning of LLMs has enjoyed great popularity in recent years, from RLHF (Christiano et al., 2017; Ouyang et al., 2022) via direct preference optimisation (DPO; Rafailov et al., 2023), to specialised workflows such as DQO (Uhlig et al., 2025). Like the present work, Berger et al. (2024) leverage machine translated segments and their corresponding post-edits as preference pairs, though they apply them to DPO on LLMs for machine translations.

The original COMET paper included a ranking-based metric version, which created a score from the Euclidean distance of the hypothesis to both the source and reference, learned via Triplet Margin Loss (Rei et al., 2020). However, this method relied on a reference, unlike our approach, and it produced scores directly from model representations rather than using a regression head. It was also not regularised, meaning that the model may produce very high and very low scores. The COMET-Rank model was dropped in later iterations of COMET.

Corpus	train	dev	test
ACED corpus	17,455	7,884	9,922
ACED preferences	2,179	1,056	959
Internal	10,000	1,000	Table 8
HELD corpus	–	–	3,000
WMT24 (z-scores)	–	–	81,589
WMT24 (prefs)	–	–	75,830
ACES	–	–	36,476

Table 2: Dataset sizes, in number of segments, used in our experiments. For details on the per-language distribution and number of test segments we create from internal data, see Appendix Table 8.

2.4 Metric Challenge Sets

ACES and SPAN-ACES (Amrhein et al., 2023; Moghe et al., 2025) are metric challenge sets intended to profile strengths and weaknesses of machine translation metrics on 68 different subcategories of translation accuracy issues. They do not cover surface issues other than some relatively easy punctuation errors. ACES covers many language pairs, though with a very small number of examples for some. It is a contrastive challenge set, meaning metrics are scored on whether the "good" translation is ranked higher than the translation with the error. The authors encourage analysing the behaviour of metrics per category, but also provide a single aggregated score ("ACES-Score"), which is a weighted sum of Kendall's τ over each of the high-level categories.

3 Datasets

In this section, we first summarise the datasets used in this study (§ 3.1), then discuss internal data quality validation via a small human annotation campaign (§ 3.2). Table 2 summarises the sizes of the datasets we use for our experiments.

3.1 Datasets Overview

Internal Localisation Data. Internally, we have a large collection of localisation-specific data across various client domains and language pairs, of which a large percentage has been post-edited by humans. We filter this data substantially, using several heuristics:

- Only use documents with any human reviews
- Only use language pairs with $> 5,000$ segments available after document filtering

- Only using segments with > 5 but < 425 characters (around the 99th percentile in length)

For this internal data, we have access to a history of edits for each segment, but no at-scale direct assessments. Despite the potential subjectivity of post-edits (cf. Popović, 2021), they are the best human-created quality signal we have. Therefore, we use preference accuracy as the main evaluation metric on our post-edited data. Details of how we transform the data into preference pairs are described in Appendix B.1

Training COMETKiwi with its standard sentence-level loss requires continuous target scores. Originally, these are normalised and re-scaled direct assessments from past WMT campaigns. However, we do not have such human data available in our setting, meaning we need to augment our data with artificial scores in order to use this loss. Details of this transformation are described in Appendix B.2.

Finally, we sample train, dev, and test splits from both types of data, ensuring that each individual client project is assigned to only one split across both datasets.

ACED Localisation Data. The ACED corpus is a public collection of English-to-German localisation data, initially for Translation Error Correction (Lin et al., 2022). For each segment, it includes a source, target, and perturbed target, where the perturbed target represents the initial hypothesis and the target is the post-edited segment. The data is spread across three sub-datasets with different domains: ASICS, (marketing copy for an activewear company), EMERSON, (industrial product listings for a manufacturer), and DIGITALOCEAN, (software engineering tutorials).

We use the predefined train, dev, and test splits, but merge them across sub-datasets for a larger data size. We apply the same transformations as for our internal data to obtain preference pairs and continuous scores. Note that the number of preference pairs is especially limited, since not all segments in ACED received a post-edit. Table 2 lists the merged data sizes for ACED.

Human-Evaluated Localisation Data (HELD). For evaluation purposes, we translate a corpus of localisation data into ten languages using a selection of different models, then collect Error Span Annotations (ESA, Kocmi et al., 2024b) from professional translators paid standard rates. We refer to this corpus as HELD from here on.

This data originates from a real client project, and includes primarily user interface text and help documentation. It contains hypotheses but no references. We normalise this data to z-scores and evaluate our quality estimation models on correlation (Kendall's τ) with the z-scores. This data is not suitable to transform into preference pairs because we only have one scored translation per segment, so we only use it to compute correlations. We aim to release this data once this article is de-anonymised.

WMT Data. The WMT conference has been publishing machine translation data with human annotations over many years. The largest domain in these datasets is news text, though since 2020 the domain coverage has been expanding: First with some Wiki data, later with more diverse domains including social media data, dialogue, e-commerce, and literary text (Barrault et al., 2020; Kocmi et al., 2024a, 2025). COMETKiwi (Rei et al., 2022) was originally trained with normalised direct assessments from the years 2017-20 (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020). We mix in data from those years during training to avoid catastrophic forgetting, and test on WMT24 to validate the general performance of our finetuned metrics. See Appendix B.3 for how we process this data.

ACES Challenge Set. We use the ACES challenge set (Amrhein et al., 2023) to validate the general capabilities of our finetuned metrics, primarily tracking the aggregated ACES-Score.

Localisation CheckList. Motivated by the observations stated in the introduction and exemplified in Table 1, we additionally create a localisation-specific challenge set addressing a set of potential issues. We think of this challenge set as a kind of CheckList (Ribeiro et al., 2020), covering behaviours that may appear simple on the surface but could get lost while optimising for other skills. The data samples are created from real localisation data, meaning its domain is aligned with our internal data. We use straightforward heuristics to create the negative samples, making the checklist reproducible. A detailed description of each heuristic can be found in Appendix D.1.

3.2 Internal Data Quality Validation

To provide a deeper sense of our internal data, we annotate a sample ($n = 984$ annotations across

System	MT	AI-PE	Human
MT	-	0.26	0.14
AI-PE	0.43	-	0.23
Human	0.61	0.52	-

Table 3: Pairwise system comparison results from an internal study, where each cell indicates the win rate of the system listed in the row when compared against the system listed in the corresponding column. MT refers to machine translation, AI-PE to automatic post-editing of MT output, and Human to final human translation.

Metric	ρ	τ
$\text{COMET}^{22}_{\text{DA}}$	0.404	0.282
$\text{COMETKiwi}^{22}_{\text{DA}}$	0.371	0.253
BLEU	0.166	0.115
chrF++	0.227	0.158
TER	-0.195	-0.137

Table 4: Segment-level Spearman R (ρ) and Kendall's Tau (τ) correlations between ESA scores and automatic evaluation metrics in the internal study. BLEU, chrF++, and TER are computed using their segment-level variants. Only MT and AI-PE annotations are included, with the final human translation serving as the reference for reference-based metrics.

317 source segments in 53 documents) of English-German hypotheses in an ESA setting (Kocmi et al., 2024b). The annotators in this small validation study are German-speaking NLP practitioners, and the data is sampled independently from HELD. The data sample is drawn randomly from our internal data, but with the criterion that there are three hypotheses present: A machine translation, an AI post-edit, and a final human post-edit. Wherever possible, we show document context.

Table 3 shows how often each type of hypothesis wins against the other hypotheses. That is, the human translation won over the MT system 61% of the time while the MT system was rated preferable over the human translation 14% of the time. The rest are ties. These win rates show how often our annotators agreed with the the edits made by the AI-PE or the human post-edit, effectively giving us a sense of how often edits are subjective.

Table 4 reports segment-level Spearman's rank correlation (ρ) and Kendall's Tau (τ) between human judgements and automatic evaluation metrics. Annotations corresponding to the final human post-edit are excluded and instead used as a reference, resulting in a total of 644 annotations. Overall, we observe slightly stronger correlations than those reported by (Rei et al., 2022). While neural metrics show very similar correlation values, surface-level metrics show substantially higher correlations compared to prior work. This can be attributed to the bias introduced by the post-editing setting in which the data was generated. Later edits tend to be close to the original version, which is measured by BLEU, TER and chrF++.

4 Metric Adaptation

4.1 Tokeniser Adjustments

Although "tokenisation" most commonly calls to mind tokeniser inference, the process of tokenisation involves multiple steps, all of which influence the outcome (cf. Schmidt et al., 2024): Normalisation, pretokenisation, tokeniser inference, and post-processing. We propose adjusting the pretokenisation step of the encoder model's tokeniser, as well as adding new tokens for special whitespace characters. These are targeted interventions which somewhat change the distribution of tokenised texts seen by the model, but leave the majority of segments unchanged. Specifically, we separate out individual digits and punctuation characters, and keep trailing and multiple spaces, which we can achieve by modifying pre-tokenisation. Additionally, we enable the model to "see" the most common special space characters: The non-breaking space, the zero-width non-joiner, and the zero-width space. Details of the tokeniser modifications are in Appendix C.

4.2 Preference Pair Fine-Tuning

We experiment with fine-tuning the metrics directly on the preference pairs induced from our post-edit data. The benefit of this approach is that the signal comes directly from human judgements, with no intervening heuristics for score creation. We use the Margin Ranking Loss (MRL; Herbrich et al., 2000) to encourage the model to score the preferred hypothesis above the rejected one:

$$\text{MRL}(x_{pos}, x_{neg}) = \max(0, -(x_{pos} - x_{neg}) + m) \quad (1)$$

where x_{pos}, x_{neg} are the scores of the preferred and rejected inputs, respectively, and the margin m is a fixed value hyperparameter. Figure 2 shows the loss schematic for MRL applied to

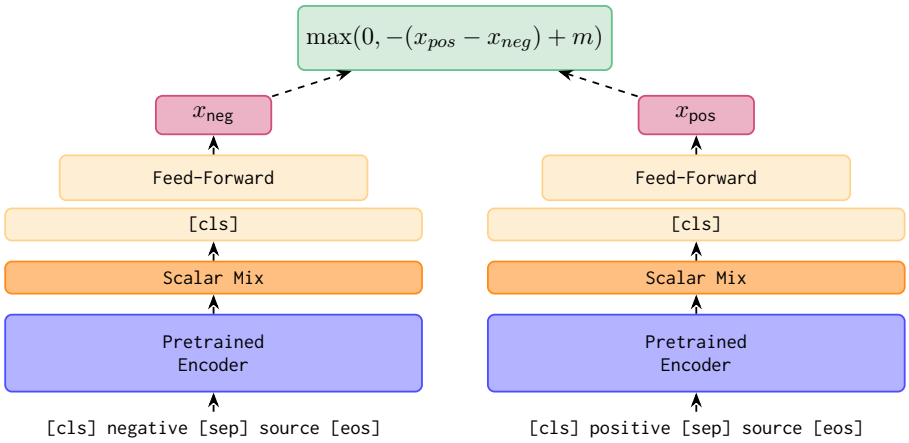

Figure 2: Margin Ranking Loss applied to a COMETKiwi model. Figure adapted from Rei et al. (2020, 2022).

COMETKiwi$^{22}_{\text{DA}}$. We set the margin to a relatively conservative 0.02 for our experiments with this loss, according to the semantics of COMET-type metrics, where a difference of 0.02 already corresponds to large difference in quality (cf. Kocmi et al., 2024c).

Additionally, we discourage the model from deviating too far from its original behaviour by using the scores of the original metric as regularisation targets. We use mean-squared error (MSE) between the predicted score x_i and the original score y_i, making the final loss:

$$\mathcal{L} = \text{MRL}(x_{pos}, x_{neg}) + \lambda_r(\text{MSE}(x_{pos}, y_{pos}) + \text{MSE}(x_{neg}, y_{neg})) \quad (2)$$

4.3 Fine-Tuning with Continuous Scores

We also fine-tune the metric models with the standard mean-squared error (MSE) loss, since this is a known working training strategy. The target scores here are derived from the chrF++ score between the translation seen by the reviewer and their post-edit (cf. Appendix B.2).

Unlike the original COMETKiwi training, we unfreeze the entire encoder, including the embedding layer, from the beginning of finetuning. Unfreezing the encoder is necessary in our case in order to train the added token embeddings. Instead of freezing the encoder, we add learning rate warmup to stabilise training, and increase both the effective batch size and the learning rates.

4.4 Fine-Tuning Settings

We fine-tune and evaluate COMETKiwi$^{22}_{\text{DA}}$ in the following settings: Fine-tuning with MRL versus MSE, with the original versus the new tokeniser, and fine-tuned on a 10k sample of our internal corpus versus the ACED corpus.

For both datasets, we mix in at least 25% data from the original COMETKiwi training sets: Due to the small size of the ACED preference data, we use a larger proportion of WMT data for contrastive training in both data mixes, reaching ca. 10k training samples each.

In addition to MSE fine-tuning with scores derived from chrF++, we also consider distilling scores from COMET$^{22}_{\text{DA}}$ into COMETKiwi. Such an approach was previously explored by Zouhar et al. (2023), and is a relevant point of comparison due to the overall higher correlation of COMET$^{22}_{\text{DA}}$ with human scores compared to chrF++.

We set the margin m for MRL to 0.02 and the regularisation weight λ_r to 0.01. We use a relatively large effective batch size of 512 samples. For 10k training samples, we use 50 warmup steps, an encoder learning rate of 1e-5, and a feedforward layer learning rate of 1e-4. The encoder, including the embeddings, is unfrozen from the beginning of training. We apply early stopping with a checkpoint every epoch, training for up to five epochs. We show individual runs with a fixed random seed.

5 Results and Discussion

Contrastive training, even with small data, improves segment-level accuracy. Figure 3 shows the accuracy of the QE models at selecting the preferred translation over the rejected on the respective target test set. The baseline COMETKiwi$^{22}_{\text{DA}}$ struggles to identify preferred segments in our internal data, while it performs better initially on the ACED corpus. Contrastive fine-tuning is effective

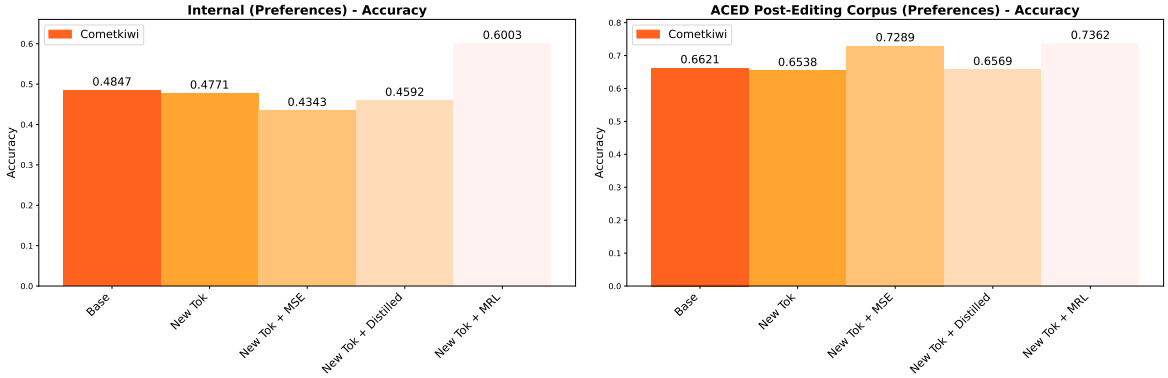

Figure 3: Performance on the respective target test set, in terms of preference accuracy. Left: Internal data. Right: ACED corpus. Per-language-pair performance for the internal data is listed in Appendix Table 9.

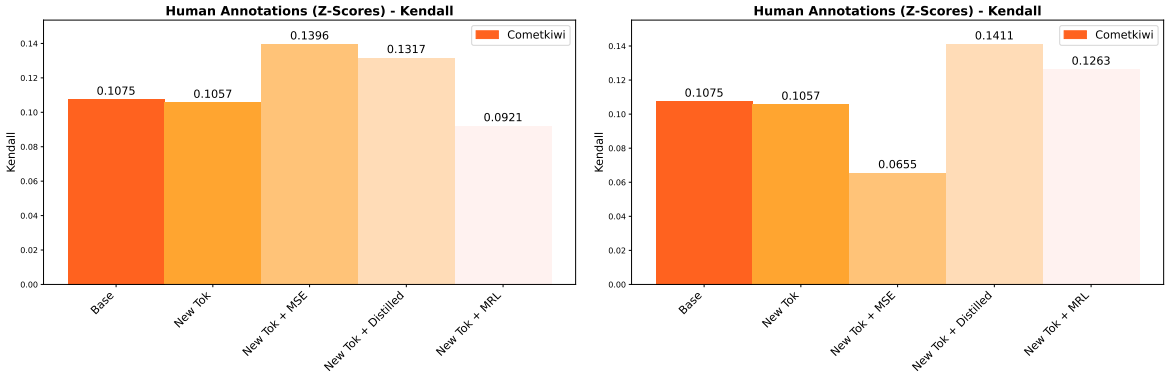

Figure 4: Overall performance on HELD, in terms of Kendall's τ. Left: Fine-tuning was done on internal data. Right: Fine-tuning was done on the ACED corpus. Per-language-pair performance is listed in Appendix Tables 10 and 11, respectively.

for improving preference accuracy in both settings, adding +11.5% and +7.5%, respectively, and allowing COMETKiwi to identify preferences on our internal data at all. MSE fine-tuning with inferred scores pushes preference accuracy below guessing on our internal data, whereas on the ACED corpus, MSE fine-tuning proves more helpful, adding up to +6.7% to accuracy. These results show that even with very small data, contrastive fine-tuning of QE models significantly improves preference accuracy on the target domain. At the same time, preference accuracy on WMT24 stays high and even improves slightly in several of the settings, as shown in Appendix Table 7, demonstrating that our training approach avoids catastrophic forgetting.

Domain adaptation with continuous scores is best for correlation. Figure 4 shows Kendall's τ of the fine-tuned QE models with z-scores from HELD. When training on our internal data, MSE fine-tuning improves Kendall by a large margin, while contrastive fine-tuning somewhat degrades performance. Meanwhile, when fine-tuning on the ACED corpus, both distillation and contrastive fine-tuning improve Kendall. Here, the best performance is achieved by the model distilled from $\text{COMET}^{22}_{\text{DA}}$, rather than trained to emulate chrF++ scores, while the second best performance comes from the model with contrastive finetuning. Distillation is effective here because $\text{COMET}^{22}_{\text{DA}}$ is known to correlate better with human ratings than chrF++, but it does not benefit preference accuracy on the ACED test set. Indeed, **correlation and segment-level accuracy conflict**, with models that perform well by one of these measures tending to do worse on the other. Even the outlier training run `New Tok + MSE` follows this pattern.

Clean data is important. Contrastive fine-tuning with the ACED corpus is the only setting that improves both the preference accuracy on the target test set and Kendall's τ on a professionally annotated dataset, while *also* keeping the ACES score high (cf. App. Table 7). Meanwhile, contrastive fine-tuning with our internal data improves preference accuracy on the internal test set while degrad-

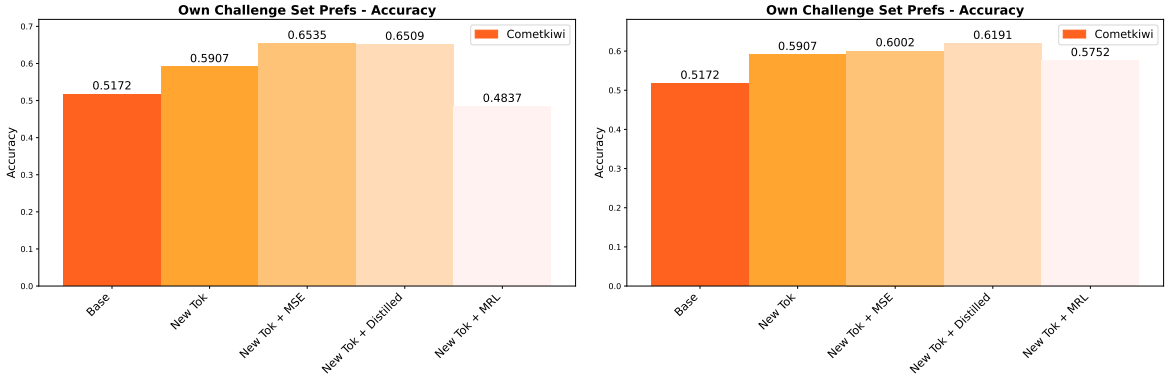

Figure 5: COMETKiwi accuracy on our Localisation CheckList. Left: Fine-tuning was done on internal data. Right: Fine-tuning was done on the ACED corpus.

ing performance on several other metrics.

The differences in behaviour when training on the ACED corpus versus the internal corpus can likely be attributed to the amount of noise in the data: Lin et al. (2022) state that not only were all translations explicitly checked by a reviewer, but they also avoided complete rewrites by replacing the original translation with the post-edit in the case of very large edit distances. The quality of the original translations in this corpus may also have been lower, making the post-edits less subjective. Additionally, the ACED corpus has a narrower scope of domains, and contains only English-to-German data. For our internal data, no manual filtering was performed, and between newer machine translation models and the AI-PE step, the human post-edits in this corpus may be more subjective (cf. § 3.2). The mix of domains and languages is also broader, making the task more challenging, and potentially introducing contradictory, domain- or language-specific preferences.

Tokeniser adaptation helps with learning specific patterns. Results on the Localisation CheckList, shown in Figure 5, demonstrate large improvements simply due to the tokeniser adjustments: +7.35%, up from near-guessing. However, as detailed in Appendix D.2, the individual CheckList categories respond differently to the new tokeniser, as well as to the two main fine-tuning approaches. As expected, the new whitespace tokens allow the model to learn whitespace rules. Contrary to our intuition, tokenising long numbers into individual digits does not help it to score them correctly. At least at our data scale, the negative effect of the longer sequence length wins out. The results further suggest that additional, targeted **data augmentation is needed to explicitly teach QE models** to attend to some of these patterns. We leave such experiments for future work.

6 Conclusions

We have demonstrated a data-efficient fine-tuning protocol for machine translation quality estimation models, adapting $\text{COMETKiwi}^{22}_{\text{DA}}$ to localisation-specific data with datasets of ca. 10k samples. To solve for a scarcity of directly scored translations in the target domain, we fine-tune models on translations and their post-edits. On one hand, we infer continuous scores using chrF++ between the machine translation and the final segment. On the other hand, we propose a regularised contrastive loss which proves particularly effective at improving segment-level ranking accuracy of the QE model. Along with targeted tokeniser interventions, our fine-tuning approaches yield QE models more suited to localisation data. Our results show trade-offs between segment-level preference accuracy and correlations with human scores.

We describe two localisation-specific datasets: The Localisation CheckList, focused on localisation-specific long-tail issues, and HELD, a human-annotated dataset of realistic translations into ten languages. We aim to release both datasets once this article is de-anonymised.

Future Work. Future work could explore additional data augmentation, particularly for the contrastive fine-tuning setting, to help QE models more reliably learn rare patterns. Further, the application of modified QE models to MT preference optimisation training, such as DQO, could potentially yield deeper insights into their behaviour.

Limitations

We acknowledge several limitations of this work. Primarily, some of the data for our main experiments is internal and cannot be published. However, we have made an effort to provide comparable experiments on similar, previously published data, showing qualitatively similar results, and to describe the internal data in detail.

The post-edits used are potentially noisy, particularly in the sense that different stylistic preferences may apply for different client projects. Our analysis in § 3.2 gives a further sense of the subjectivity of edits, indicating that not every preference pair is likely to yield a helpful signal. Nevertheless, our results show promising results for domain adaptation even with noisy data.

Additionally, we have demonstrated our approach on only one type of quality estimation model, although it may in principle apply to other models as well. This was done mainly for efficiency reasons, as COMETKiwi is a relatively lightweight model that makes it easier to inference and fine-tune.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nathaniel Berger, Miriam Exel, Matthias Huck, and Stefan Riezler. 2024. Post-edits are preferences too. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1289–1300, Miami, Florida, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin bank boundaries for ordinal regression. In *Advances in Large-Margin Classifiers*, pages 115–132. MIT Press.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models.

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, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. 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.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024c. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. 2022. Automatic correction of human translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–507, Seattle, United States. Association for Computational Linguistics.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, Frédéric Blain, Eva Vanmassenhove, Mirella De Sisto, Chris Emmery, and Pieter Spronck. 2023. Tailoring domain adaptation for machine translation quality estimation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 9–20, Tampere, Finland. European Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Kaden Uhlig, Joern Wuebker, Raphael Reinauer, and John Denero. 2025. Cross-lingual human-preference alignment for neural machine translation with direct quality optimization. In *Proceedings of the Tenth Conference on Machine Translation*, pages 31–51, Suzhou, China. Association for Computational Linguistics.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. Poor man's quality estimation: Predicting reference-based MT metrics without the reference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325, Dubrovnik, Croatia. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

A AI Assistant Information

Coding assistants were used in implementing the data processing, fine-tuning approaches, and creating the plots. The outputs were thoroughly checked and tested for purpose. No LLMs were used in the writing of this article.

B Data Processing Details

B.1 Preference Pairs from Post-Edits

To create the preference pairs from our post-edit data, we filter for segments where the final translation was provided or approved by a human reviewer, treating this as a preferred translation. As negative samples, we use previous machine-translated versions of the segment, where they differ from the final version. To obtain a reference, we filter further for segments which have more than one human-reviewed version, taking the final translation as the reference and the first human-reviewed version as the preferred. This second filtering step restricts the data much further, but still yields plenty of data across 51 language pairs. Table 8 lists absolute test set sizes, and relative train and development set sizes across all language pairs used.

Due to the strict requirement of using a human-edited segment as the preferred translation, we can be confident that the preferred is considered better than the previous version by the respective client. Similarly, we are confident that the reference is a good translation for the given segment. Regardless of whether the first human-reviewed version still has flaws, or the further edit is merely for subjective style, the preferred translation is better than the negative sample. A limitation of these references is that subsequent edits in a second review step are typically smaller than the first post-editing step, since the largest errors were already caught. Therefore, the reference will often be very similar to the preferred segment.

We apply the same approach to the ACED corpus, but in this setting we only have access to two versions of the translation, of which one is the preferred. Since we do not want to use identical preferred and reference segments, we can use this preference data only without a reference.

B.2 Continuous Scores from Post-Edits

Using HTER (human-targeted translation edit rate), as part of the training signal for metrics is an established practice (e.g., Sharami et al., 2023). That said, TER is somewhat unstable on a segment level,

especially when segments are short—which is a common occurrence in our data. Thus, we select chrF++ as our surface metric for measuring post-edit distance. Additionally, we use heuristics to identify numbers, URLs, and "DNTs" (**D**o **N**ot **T**ranslate) which the reviewer had to edit. Since these are semantically major errors which may not result in a large edit distance, we subtract an additional 20 chrF points for these cases. Finally, we calculate z-scores from these artificial scores, and rescale them to between [0,1].

B.3 WMT Data Processing

To use the WMT data in both our training schemes, we need both continuous scores and preference pairs. As continuous scores, we naturally use the z-scores from the DA/SQM or ESA annotations, rescaled to between [0,1]. To transform the data into preference pairs, we compare both the raw scores and z-scores of all submissions for a given segment. If the z-scores of two submissions are more than one standard deviation apart, we turn them into a preference pair. The raw scores act as a sanity check: We never use a submission as the rejected translation if its raw score is 98 or higher.

C Details of Tokeniser Modifications

The pretrained encoder used by $\text{COMETKiwi}^{22}_{\text{DA}}$ is based on XLM-R (Conneau et al., 2020), specifically, InfoXLM (Chi et al., 2021). Thus, it uses the XLM-R tokeniser, with a vocabulary of ca. 250k subword tokens, trained with SentencePiece (Kudo and Richardson, 2018) and implemented for inference in the Huggingface `tokenizers` library.

By default, the XLM-R tokeniser splits on any whitespace, stripping away trailing and multiple spaces. It also replaces spaces with the "metaspace" character. Since we are only inferencing the tokeniser, it has already learned exclusively tokens that do not cross whitespace boundaries. Rather than implement a new whitespace pretokeniser, we can simply remove this processing step to preserve trailing and multiple spaces without significantly changing the subword token sequence, as long as we keep the metaspace replacement.

We add pretokeniser steps to separate out digits and punctuation characters, meaning that any subword tokens consisting of multiple digits, multiple punctuation characters, or a mixture of these, can no longer be produced. Since these represent only around 1.3% of the embedding layer, we do not take additional steps to remove them from the model at this stage.

Regarding the special space characters, the `tokenizers` library allows us to add tokens which bypass the Unicode normalisation step. We add the non-breaking space (NBSP), the zero-width non-joiner (ZWNJ), and the zero-width space (ZWSP) as additional tokens. In the model's embedding layer, the new tokens are initialised according to the AvgEmb procedure shown by Hewitt (2021).

D Localisation CheckList Details

D.1 Composition

We include eleven language pairs in our Localisation CheckList, all translating from English into the following target languages: Arabic, German, Spanish, Finnish, French, Hindi, Japanese, Polish, Russian, Thai, and Chinese. These target languages cover both very high-resource and more mid-resource language pairs, seven different scripts, and as many language families. Each sample in the challenge set consists of a source segment, a reference segment, a positive sample (a "good" translation that meets the criterion to be tested), and a negative sample which is identical to the positive sample except for the variable under study. We ensure that the reference differs from the positive sample by using the final translation as the reference, and an intermediate translation that meets the criterion as the positive sample. We apply the following heuristics to find and create positive and negative samples:

- **Copy URLs.** Although URLs may be localised, an MT model cannot know the localised URL and so should be expected to copy the source URL if no other instruction is given. We sample translations where a URL was changed during the post-editing process. We ensure that the reference and positive sample contain the source URL, then place the changed URL in the negative sample.

- **Copy DNTs. D**o **N**ot **T**ranslate items are elements within the segment, usually marked by a specific syntax, that should be copied to the correct place in the target segment, and must not be translated. We sample segments that organically contain one or more DNTs, and prompt Gemma3-27B-IT to provide a translation for one of the DNTs.

	en-ar	en-de	en-es	en-fi	en-fr	en-hi	en-ja	en-pl	en-ru	en-th	en-zh	all
Copy URLs	8	29	50	5	50	1	50	10	23	9	44	279
Copy DNTs	–	1	7	8	4	14	12	2	11	3	1	63
Long Numbers	50	50	50	50	50	50	50	50	50	34	50	534
Lead./Trail. Sp.	50	50	50	50	50	50	50	50	50	50	50	550
Extra Period	50	50	50	50	50	50	50	50	50	–	50	500
All Caps	–	50	50	50	50	–	–	50	50	–	–	300
Spanish ¿ & ¡	–	–	50	–	–	–	–	–	–	–	–	50
French NBSPs	–	–	–	–	50	–	–	–	–	–	–	50
Totals	158	230	307	213	304	165	212	212	234	96	195	**2,326**

Table 5: Distribution of samples in the Localisation CheckList challenge set.

CheckList Category	$\text{COMETKiwi}^{22}_{\text{DA}}$	+New Tok	+MSE	+Tok+MSE	+MRL	+Tok+MRL
Copy URLs	82.7	79.8	**89.9**	87.9	52.4	52.0
Copy DNTs	63.2	50.9	**94.7**	82.5	50.9	43.9
Long Numbers	85.2	80.5	**96.3**	87.6	44.0	45.7
Lead./Trail. Sp.	24.9	**66.0**	26.9	51.5	18.7	27.1
Extra Period	40.2	42.4	40.6	43.4	**50.2**	41.4
All Caps	22.3	25.7	43.0	39.0	48.7	**50.7**
Spanish ¿ & ¡	**96.0**	48.0	62.0	66.0	**96.0**	**96.0**
French NBSPs	44.0	18.0	44.0	44.0	60.0	**82.0**

Table 6: Performance of COMETKiwi models on the Localisation CheckList set by error category. Fine-tuning was done with the internal dataset throughout.

- **Long Numbers.** If a number is mistranslated, the metric should catch this and consider it a major error, even if the number is long. We sample translations organically containing long digit sequences, and check that the intermediate translation matches the reference. Then, we randomly remove, add, or swap one or more digits, and place the resulting digit sequence within the negative sample.

- **Leading/Trailing Spaces.** Translators are careful to leave leading and trailing spaces in a segment unchanged, but the default tokenisation of COMET models drops these entirely. We sample translations with leading or trailing spaces, then randomly add, remove, or replace whitespace characters from these leading or trailing space sequences. The whitespace characters we sample from for addition or replacement are the regular space, tab, NBSP, ZWSP, and ZWNJ.

- **Extra Period.** If a longer segment does not end in a period, language models still sometimes prefer a translation ending in a period, but translators do not. We sample segments with a minimum of ten source words that do not end in a period, then add the appropriate sentence-ending symbol for each script to the negative sample.

- **All Caps.** If the source segment is written in all-caps, a translation in all-caps should be preferred to a translation in sentence case. We sample segments in sentence case and use the initial translation as the negative sample, then upper-case the source, positive sample, and reference. This criterion is only applicable in Latin and Cyrillic script.

- **Spanish ¿ and ¡.** We want to verify that the sentence-initial exclamation mark and question mark in Spanish are treated correctly by the metrics. We sample Spanish sentences containing these characters and remove them to create the negative sample.

- **French NBSPs.** French has specific rules about where non-breaking spaces should be used, but COMET models cannot differentiate special space characters. We sample French segments where a NBSP occurs with punctuation, and randomly either remove the NBSP or replace it with a regular whitespace character.

Table 5 shows the composition of the Localisation CheckList by language pair and error category.

D.2 Analysis of Results

Overall results are visualised in Figure 5. Table 6 breaks down the performance of COMETKiwi variants, fine-tuned on internal data, by category.

URLs, DNTs, and long numbers show a strong improvement under MSE fine-tuning. Recall that we explicitly subtracted from the chrF scores assigned to segments failing these criteria, so the QE model has clearly learned to attend to these patterns due to this data augmentation. Changing the tokeniser harms these categories significantly, indicating that these segments are rather strongly affected by longer sequence lengths, and that at least with small training data, seeing individual digits does not help the QE model to score number translations correctly. Contrastive fine-tuning unfortunately harms these categories.

As expected, the category of leading and trailing spaces is strongly helped by allowing the model's tokeniser to see the whitespace in question. Fine-tuning harms performance here, likely not providing a useful signal for this category. Regarding the NBSP, merely adding it to the tokeniser harms performance because there is now a new, untrained token in many of the segments. Fine-tuning with MSE restores performance to baseline—chrF++ does not distinguish the NBSP either. Here, however, contrastive fine-tuning proves effective: Our localisation data clearly contains a number of segments where this type of edit was made, and consistently so.

Accuracy on the 'Extra Period' category stays very similar, again suggesting a lack of useful signal. Additionally, contrastive fine-tuning is somewhat more effective at making the QE model score an all-caps translation of an all-caps source above the sentence-case version, though many segments also remain unchanged.

The Spanish punctuation marks are an outlier: The base QE model shows near-perfect performance on this category, but splitting punctuation marks into their own tokens puts it at guessing level. Fine-tuning with MSE restores some performance compared to only changing the tokeniser, but hurts compared to the baseline. Meanwhile, contrastive fine-tuning fully restores performance to baseline.

E Additional Figures and Tables

Table 7 shows the ACES-Score (Amrhein et al., 2023) achieved by fine-tuned metrics, a proxy for their general capabilities. It also lists their preference accuracy on the WMT24 test set. Table 8 lists the dataset sizes of our internal data per language pair, while Table 9 shows per-language-pair performance of fine-tuned metric and QE models on that same data. Tables 10 and 11 break down the per-language-pair performance of our QE models on HELD.

Model	ACES	WMT24 Acc.
ACED fine-tuning		
Cometkiwi (Base)	**17.94**	65.5
+ New Tok	17.45	65.6
+ MSE	14.79	66.2
+ New Tok + MSE	12.87	64.8
+ New Tok + Distilled	12.80	65.9
+ MRL	17.22	**66.9**
+ New Tok + MRL	16.95	**66.9**
Internal data fine-tuning		
+ MSE	16.86	65.3
+ New Tok + MSE	16.49	65.1
+ New Tok + Distilled	16.54	65.8
+ MRL	13.09	**67.5**
+ New Tok + MRL	13.27	67.2

Table 7: General metric capabilities, in terms of ACES-Score and preference accuracy (in percent) on WMT24.

Dataset	de-en	de-es	de-fr	en-ar	en-bg	en-bn	en-cs	en-da	en-de
Train (Inferred Scores)	0.48%	0.20%	0.21%	3.58%	0.70%	0.27%	1.93%	2.76%	4.60%
Train (Preferences)	0.25%	0.05%	0.04%	2.39%	0.34%	0.14%	1.96%	1.66%	7.41%
Dev (Inferred Scores)	0.57%	–	–	1.55%	0.04%	0.13%	0.55%	1.70%	2.84%
Dev (Preferences)	0.33%	–	–	0.93%	0.03%	0.07%	0.46%	1.33%	4.20%
Test (Inferred Scores)	269	500	500	500	500	500	444	500	500
Test (Preferences)	19	40	18	500	18	65	70	268	500
Dataset	**en-el**	**en-es**	**en-fa**	**en-fi**	**en-fr**	**en-he**	**en-hi**	**en-hr**	**en-hu**
Train (Inferred Scores)	1.10%	5.09%	0.32%	5.66%	4.57%	0.47%	1.77%	0.84%	0.57%
Train (Preferences)	0.65%	7.72%	0.10%	5.35%	6.97%	0.32%	1.36%	0.77%	0.49%
Dev (Inferred Scores)	2.04%	1.90%	0.50%	0.09%	2.51%	2.32%	1.47%	0.19%	1.89%
Dev (Preferences)	1.10%	4.13%	0.31%	0.06%	3.69%	1.41%	1.20%	0.16%	2.31%
Test (Inferred Scores)	361	500	–	252	500	500	500	500	500
Test (Preferences)	21	500	–	19	500	105	98	139	466
Dataset	**en-hy**	**en-id**	**en-it**	**en-ja**	**en-ko**	**en-lt**	**en-lv**	**en-mr**	**en-ms**
Train (Inferred Scores)	0.10%	1.85%	4.04%	4.40%	3.97%	0.26%	0.25%	0.19%	1.29%
Train (Preferences)	0.04%	1.19%	4.23%	7.27%	3.71%	0.14%	0.13%	0.08%	0.27%
Dev (Inferred Scores)	–	1.74%	9.11%	6.77%	7.57%	0.43%	0.01%	1.25%	–
Dev (Preferences)	–	1.14%	9.85%	7.71%	7.89%	0.34%	0.01%	0.34%	–
Test (Inferred Scores)	500	500	500	500	500	500	490	–	500
Test (Preferences)	138	500	500	500	500	42	34	–	416
Dataset	**en-nl**	**en-no**	**en-pl**	**en-pt**	**en-ro**	**en-ru**	**en-sk**	**en-sl**	**en-so**
Train (Inferred Scores)	4.78%	2.97%	4.38%	4.23%	2.18%	4.18%	0.46%	0.31%	0.25%
Train (Preferences)	4.86%	2.37%	2.99%	6.82%	1.87%	3.09%	0.30%	0.16%	0.22%
Dev (Inferred Scores)	5.29%	1.58%	4.12%	3.43%	5.64%	0.85%	1.34%	–	–
Dev (Preferences)	5.11%	0.97%	1.86%	6.12%	5.48%	0.72%	0.73%	–	–
Test (Inferred Scores)	500	500	500	500	500	500	500	500	500
Test (Preferences)	500	95	500	500	39	398	7	28	53
Dataset	**en-sq**	**en-sv**	**en-sw**	**en-ta**	**en-te**	**en-th**	**en-tl**	**en-tr**	**en-uk**
Train (Inferred Scores)	0.15%	2.49%	0.01%	0.18%	0.25%	3.84%	0.86%	1.56%	3.74%
Train (Preferences)	0.10%	1.10%	0.01%	0.10%	0.20%	2.55%	0.48%	1.25%	4.19%
Dev (Inferred Scores)	–	7.89%	–	0.67%	–	4.32%	–	6.70%	0.55%
Dev (Preferences)	–	4.90%	–	0.25%	–	5.65%	–	6.38%	0.68%
Test (Inferred Scores)	500	500	500	–	–	500	500	500	500
Test (Preferences)	20	369	173	–	–	195	4	98	84
Dataset	**en-ur**	**en-vi**	**en-zh-Hans**	**en-zh-Hant**	**es-en**	**pt-en**			**Total**
Train (Inferred Scores)	0.16%	2.07%	4.35%	3.71%	1.25%	0.14%			100.0%
Train (Preferences)	0.10%	1.39%	6.41%	3.56%	0.75%	0.10%			100.0%
Dev (Inferred Scores)	0.74%	0.34%	3.65%	2.69%	2.87%	0.14%			100.0%
Dev (Preferences)	0.46%	0.17%	6.49%	3.59%	1.40%	0.04%			100.0%
Test (Inferred Scores)	–	500	500	500	–	500			21,816
Test (Preferences)	–	251	500	500	–	202			10,492

Table 8: Internal dataset sizes by language pair.

Model	en-ar	en-bn	en-cs	en-da	en-de	en-es	en-fr	en-he	en-hi
Cometkiwi (Base)	0.584	**0.800**	0.486	0.410	0.414	0.462	0.428	0.514	0.541
+ New Tok	0.586	0.738	0.371	0.444	0.380	0.472	0.398	0.505	0.490
+ Finetuned (MSE)	0.506	0.769	0.414	0.444	0.390	0.420	0.372	**0.581**	0.531
+ New Tok + Finetuned (MSE)	0.496	0.738	0.343	0.407	0.376	0.422	0.372	0.571	0.469
+ New Tok + Distilled (MSE)	0.552	0.723	0.414	0.392	0.404	0.438	0.392	0.552	0.520
+ Finetuned (MRL)	0.602	0.723	**0.729**	0.493	**0.586**	0.520	0.552	0.552	**0.571**
+ New Tok + Finetuned (MRL)	**0.616**	**0.800**	0.714	**0.504**	0.578	**0.528**	**0.558**	0.543	0.541
Model	**en-hr**	**en-hu**	**en-hy**	**en-id**	**en-it**	**en-ja**	**en-ko**	**en-ms**	**en-nl**
Cometkiwi (Base)	0.568	0.502	0.478	0.500	0.514	0.492	0.550	0.399	0.468
+ New Tok	0.590	0.509	0.500	0.490	0.514	0.478	0.550	0.406	0.460
+ Finetuned (MSE)	0.511	0.491	0.507	0.486	0.466	0.464	0.522	0.298	0.430
+ New Tok + Finetuned (MSE)	0.525	0.491	0.522	0.486	0.456	0.458	0.520	0.305	0.416
+ New Tok + Distilled (MSE)	0.568	0.491	0.536	0.534	0.486	0.494	0.530	0.327	0.444
+ Finetuned (MRL)	**0.604**	0.639	**0.601**	**0.588**	**0.636**	**0.676**	**0.662**	0.591	0.632
+ New Tok + Finetuned (MRL)	**0.604**	**0.657**	0.594	0.584	0.632	0.662	0.644	**0.625**	**0.646**
Model	**en-no**	**en-pl**	**en-pt**	**en-ru**	**en-so**	**en-sv**	**en-sw**	**en-th**	**en-tr**
Cometkiwi (Base)	0.600	0.344	0.486	0.465	0.302	0.439	0.566	0.610	0.704
+ New Tok	0.600	0.336	0.482	0.430	0.358	0.420	0.572	0.590	0.684
+ Finetuned (MSE)	0.579	0.348	0.432	0.397	0.264	0.415	0.543	0.528	0.643
+ New Tok + Finetuned (MSE)	0.579	0.304	0.414	0.387	0.302	0.396	0.584	0.564	0.643
+ New Tok + Distilled (MSE)	0.600	0.318	0.432	0.422	0.283	0.434	0.595	0.605	0.673
+ Finetuned (MRL)	**0.684**	**0.560**	**0.588**	**0.633**	0.358	**0.515**	**0.653**	**0.697**	0.714
+ New Tok + Finetuned (MRL)	0.663	0.550	0.576	0.601	**0.396**	0.512	**0.653**	0.646	**0.765**
Model	**en-uk**	**en-vi**	**en-zh-Hans**	**en-zh-Hant**	**pt-en**	**xx-yy**			
Cometkiwi (Base)	0.631	0.602	0.486	0.492	0.302	0.485			
+ New Tok	0.619	0.586	0.476	0.492	0.302	0.477			
+ Finetuned (MSE)	0.357	0.510	0.430	0.420	0.213	0.442			
+ New Tok + Finetuned (MSE)	0.345	0.514	0.430	0.418	0.213	0.434			
+ New Tok + Distilled (MSE)	0.393	0.570	0.464	0.428	0.243	0.459			
+ Finetuned (MRL)	**0.702**	0.713	0.568	0.610	**0.455**	0.600			
+ New Tok + Finetuned (MRL)	0.679	**0.721**	**0.588**	**0.626**	**0.455**	**0.600**			

Table 9: (Finetuning on internal data) Preference accuracy on internal data, by language pair. Only listing language pair splits with at least 50 preference pair samples in the respective test set.

Model	en-ar	en-de	en-es	en-fi	en-fr	en-ja	en-pl	en-pt	en-sv	en-tr
Cometkiwi (Base)	0.175	0.119	0.103	0.133	0.173	0.164	0.136	0.086	0.110	0.235
+ New Tok	0.165	0.120	0.082	0.131	0.153	0.142	**0.157**	0.093	0.103	0.230
+ Finetuned (MSE)	0.175	**0.140**	0.135	**0.170**	**0.208**	**0.166**	0.126	0.153	**0.151**	0.232
+ New Tok + Finetuned (MSE)	**0.183**	0.128	**0.138**	0.166	0.177	0.156	0.124	**0.172**	0.146	0.216
+ New Tok + Distilled (MSE)	0.166	0.122	0.127	0.166	0.172	0.142	0.126	0.144	0.133	0.235
+ Finetuned (MRL)	0.173	0.101	0.098	0.127	0.181	0.160	0.114	0.091	0.075	**0.237**
+ New Tok + Finetuned (MRL)	0.165	0.092	0.107	0.106	0.167	0.163	0.106	0.086	0.072	0.233

Table 10: (Finetuning on internal data) Kendall's τ of QE models with human annotations (z-scores), by language pair.

Model	en-ar	en-de	en-es	en-fi	en-fr	en-ja	en-pl	en-pt	en-sv	en-tr
Cometkiwi (Base)	0.175	0.119	0.103	0.133	0.173	0.164	0.136	0.086	0.110	0.235
+ New Tok	0.165	0.120	0.082	0.131	0.153	0.142	0.157	0.093	0.103	0.230
+ Finetuned (MSE)	0.151	0.119	0.113	0.162	0.189	0.175	0.104	0.103	0.125	**0.269**
+ New Tok + Finetuned (MSE)	0.138	0.109	0.081	0.115	0.120	0.137	0.076	0.057	0.118	0.230
+ New Tok + Distilled (MSE)	0.176	**0.177**	**0.157**	**0.242**	0.192	**0.206**	**0.186**	**0.171**	**0.143**	0.262
+ Finetuned (MRL)	**0.181**	0.126	0.125	0.154	**0.242**	0.190	0.127	0.118	0.105	0.252
+ New Tok + Finetuned (MRL)	0.175	0.125	0.111	0.137	0.204	0.184	0.140	0.123	0.108	0.255

Table 11: (Finetuning on ACED) Kendall's τ of QE models with human annotations (z-scores), by language pair.