Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains

Vilém Zouhar[1*]   **Shuoyang Ding**[2]   **Anna Currey**[2]
Tatyana Badeka[2]   **Jenyuan Wang**[2]   **Brian Thompson**[2†]
[1]ETH Zürich   [2]AWS AI Labs
brianjt@amazon.com

Abstract

We introduce a new, extensive multidimensional quality metrics (MQM) annotated dataset covering 11 language pairs in the biomedical domain. We use this dataset to investigate whether machine translation (MT) metrics which are fine-tuned on human-generated MT quality judgements are robust to domain shifts between training and inference. We find that fine-tuned metrics exhibit a substantial performance drop in the unseen domain scenario relative to both metrics that rely on the surface form and pre-trained metrics that are not fine-tuned on MT quality judgments.

1 Introduction

Automatic metrics are vital for machine translation (MT) research: given the cost and effort required for manual evaluation, automatic metrics are useful for model development and reproducible comparison between research papers (Ma et al., 2019). In recent years, the MT field has been moving away from string-matching metrics like BLEU (Papineni et al., 2002) towards fine-tuned metrics like COMET (Rei et al., 2020), which start with pre-trained models and then fine-tune them on human-generated quality judgments. Fine-tuned metrics have been the best performers in recent WMT metrics shared task evaluations (Freitag et al., 2022, 2023) and are recommended by the shared task organizers, who go so far as to say, *"Neural fine-tuned metrics are not only better, but also robust to different domains."* (Freitag et al., 2022).

Given the growing popularity of fine-tuned metrics, it is important to better understand their behavior. Here, we examine the question of domain robustness of fine-tuned metrics. Fine-tuned metrics contain extra parameters on top of the pre-trained model which are initialized randomly (or to zero) and then fine-tuned on human-generated MT
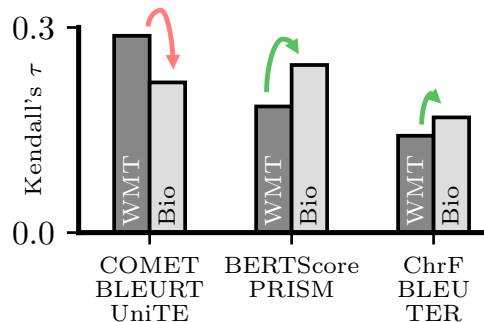

Figure 1: Automatic machine translation metric performance on the WMT and biomedical domains, averaged across metric types (see Figure 2 for full results).

quality annotations. The primary source of those annotations is prior WMT metrics shared tasks, and domains in WMT are often carried over from year to year (e.g. news). This raises the question: are fine-tuned metrics in fact robust across any domain (including domains not seen in training)? Or can their apparent strong performance be attributed in part to the artificially good domain match between training and test data?

To answer these questions, we first collect human multidimensional quality metrics (MQM) annotations in the biomedical (bio) domain. Vocabulary overlap and error analysis suggest that this new dataset is distinct from the domains used in WMT. This data covers 11 language pairs and 21 translation systems, with 25k total judgments. In addition to the MQM annotations, we also create new high-quality reference translations for all directions. We release this data publicly, along with code for replication of our experiments.[1]

Next, we examine how different types of metrics perform on our new bio test set relative to the WMT test set. We find that fine-tuned metrics have substantially lower correlation with human

---

*Work done during an internship at Amazon.

†Corresponding author

[1]github.com/amazon-science/bio-mqm-dataset

Architecture	Metrics
Surface-Form $\frac{tgt}{ref}$ → ⚙Metric → score	BLEU CHRF TER
Pre-trained+Algorithm $\frac{src}{tgt}$ref → 🧠Model ⚙Metric → score	BERTSCORE PRISM
Pre-trained+Fine-tuned $\frac{src}{tgt}$ref → 🧠LLM 🧠Metric → score	COMET UNITE BLEURT
Pre-trained+Prompt $\frac{src}{tgt}$ref → 🧠LLM 🔮Metric → score	GEMBA AUTOMQM

Table 1: Metric types considered in this work. The 🧠 components have trainable parameters while ⚙ use handcrafted heuristics or algorithms and 🔮 decodes from a language model. The *ref* input is omitted in the case of reference-free metrics (i.e. quality estimation).

		WMT	Bio
Error severity	Critical	N/A	8%
	Major	26%	44%
	Minor	43%	31%
	Neutral	31%	16%
Error category	Fluency	47%	66%
	Accuracy	44%	18%
	Terminology	6%	10%
	Locale	2%	2%
	Other	1%	4%
Error-free segments		45%	72%
Errors per erroneous segment		1.9	2.1
Abs. erroneous segment score		-4.1	-7.6

Table 2: Error distribution of our new bio dataset and the existing WMT22 MQM dataset. The MQM annotation scheme for WMT in most cases did not contain the *Critical* category.

judgments in the bio domain, despite other types of metrics having higher correlation in the bio domain (see Figure 1), indicating they struggle with the training/inference domain mismatch. Finally, we present analysis showing that this performance gap persists throughout different stages of the fine-tuning process and is not the result of a deficiency with the pre-trained model.

2 Related Work

Metric types. Table 1 summarizes the different types of metrics that are commonly used to evaluate MT. The earliest type of MT metrics are *Surface-Form* metrics, which are purely heuristic and use word- or character-based features. We consider three common *Surface-Form* metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and CHRF (Popović, 2015). Metrics like COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and UNITE (Wan et al., 2022) start with a pretrained language model and fine-tune it on human-generated MT quality judgments. We denote these metrics *Pre-trained+Fine-tuned*.[2] Another class of metrics also start with a pre-trained model but do not perform fine-tuning. Examples of such metrics include PRISM (Thompson and Post, 2020a,b), which uses the perplexity of a neural paraphraser, and BERTSCORE (Sun et al., 2022), which is based on cosine similarity of word embeddings. We denote such metrics *Pre-trained+Algorithm* metrics. More recently, metrics like GEMBA

(Kocmi and Federmann, 2023) and AUTOMQM (Fernandes et al., 2023) have proposed prompting a large language model. We denote these as *Pre-trained+Prompt* metrics.

Domain specificity. Domain specificity for MT metrics was first explored by C. de Souza et al. (2014) for *Surface-Form* metrics. Sharami et al. (2023) brought attention to the issue of domain adaptation for quality estimation (QE), offering solutions based on curriculum learning and generating synthetic scores similar to Heo et al. (2021), Baek et al. (2020), and Zouhar et al. (2023). Sun et al. (2022) examined general-purpose natural language generation metrics and documented their bias with respect to social fairness. For word-level QE, Sharami et al. (2023) reported the lack of robustness of neural metrics.

3 New Bio MQM Dataset

We create and release new translations and MQM annotations for the system submissions from 21 participants to the WMT21 biomedical translation shared task (Yeganova et al., 2021). To explore how different the bio domain is from the WMT22 metric task domains, we computed the vocabulary overlap coefficient between each domain. Bio had the smallest average overlap with the WMT domains (0.436) compared to 0.507, 0.486, 0.507, and 0.582 for e-commerce, news, social, and conversation, respectively. See Appendix A for full details and example sentences from each domain.
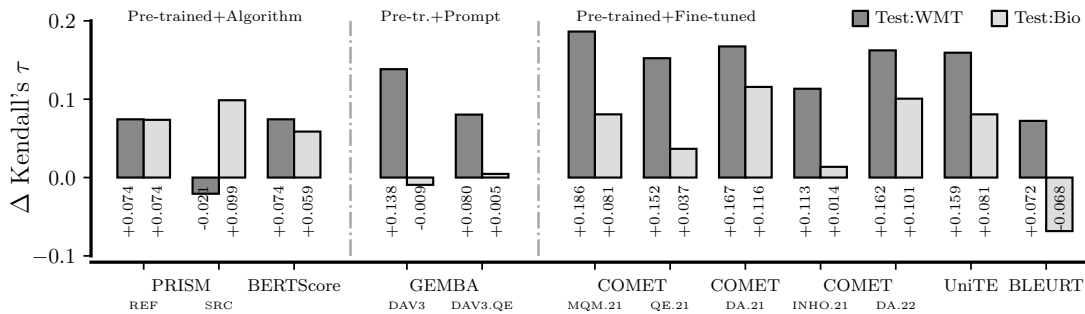
---

[2]The WMT metrics task calls these "trained" metrics.

Figure 2: Gains in segment-level correlation (Kendall's τ) when comparing *Surface-Form* metrics (average performance of BLEU, CHRF, and TER) to a given metric, on the WMT and bio test sets. Gains for *Pre-trained+Fine-tuned* metrics are much smaller in the unseen bio domain than the WMT domain. *Pre-trained+Algorithm* metrics, which do not train on prior WMT data, do not exhibit the same bias. See Appendix F for results in tabular form.

3.1 Dataset Creation

We created the bio MQM dataset in three steps. Annotations and translations were performed by expert linguists with experience in the medical domain (see Appendix C for full details).

Step 1: Reference re-translation. The original bio test set consists of bilingual abstracts from crawled academic papers, which might be written by non-native speakers (Névéol et al., 2020) or even MT (Thompson et al., 2024). Therefore, we create new professional reference translations.

Step 2: Reference quality. To ensure a high bar of quality for the reference translations, we ask a separate set of annotators to provide MQM annotations for the new references. Any issues identified by this round of MQM annotation are then fixed by a new set of translators, resulting in the final reference translations that we release in this dataset.

Step 3: MQM annotations. Finally, we conduct the main MQM annotation on the references and shared task system outputs. In this step, a single annotator rates all translations of a given document (from all systems and the reference).[3] Our MQM schema follows Freitag et al. (2021) except that we add a *Critical* severity (assigned the same score as *Major* for backward compatibility). Full annotator instructions are in Appendix D.

The resulting dataset contains roughly 25k segment-level annotations spanning 11 translation directions.[4] In contrast, most publicly available MQM data to date covers only a few language pairs.

We use ~25% of the segments for each language pair as the train/dev set, leaving the rest as the test set (see Appendix B for exact sizes in each pair).

We compare error distributions on our new bio MQM dataset and the existing WMT MQM dataset in Table 2. Bio MQM contains more *Critical/Major* errors, and lower absolute scores on average. However, WMT MQM has more overall sentences where an error occurs. Error category distribution also diverges, notably in *Fluency* and *Accuracy*.

4 Analysis

4.1 Are fine-tuned metrics robust across domains?

Measuring domain robustness. The performance of a MT metric is typically measured by a certain *meta-evaluation metric*, such as segment-level Kendall's τ correlation with human judgments. Intuitively, one could simply measure domain robustness by comparing the performance of a certain metric on domain A and domain B. This, however, is not straightforward with meta-evaluations for metrics, since performance measured by those meta-evaluations is also affected by factors such as the quantity and quality of the translations included in the dataset, which is often hard to control for.

As a result, we resort to comparisons of *relative* performance measured against a domain-invariant baseline. To establish such comparison, we make two assumptions:

1. We assume *Surface-Form* metrics can serve as a domain-invariant baseline, as they are purely based on heuristics and do not involve parameters specifically tuned on a certain domain. We use average performance of BLEU,

---

[3]This allows us to distribute annotation jobs to multiple annotators while still allowing the annotator to access document-level context and ensuring that the whole document is ranked consistently.

[4]Pt→En, En↔De, En↔Es, En↔Ru, En↔Fr, Zh↔En

CHRF, and TER as the baseline to minimize the impact of specific choice of heuristics.

2. We assume segment-level Kendall's τ correlation with human judgments has a linear relationship with the objective performance of a metric. Hence, relative performance can be measured by simple linear subtraction.

Observations. Compared to *Surface-Form* metrics, we find that *Pre-trained+Fine-tuned* metrics provide a substantially smaller (sometimes even negative) improvement in human correlation in the bio domain than the WMT domain (see Figure 2). On the other hand, *Pre-trained+Algorithm* metrics, which have not been trained on WMT data, do not exhibit the same gap. This gap suggests that fine-tuned metrics struggle with unseen domains.

We also observe a very large performance gap for *Pre-trained+Prompt* metrics. Unfortunately, these metrics rely on closed-source LLMs without published training procedures, so we do not know what data the underlying LLMs were trained on.

4.2 How does fine-tuning affect domain robustness?

Model description. For this section, we focus on COMET (reference-based) and COMET-QE (reference-free) as they are among the most commonly used MT metrics. The COMET model works by representing the source, the hypothesis and the reference as three fixed-width vectors using a language model, such as XLM-Roberta-large (Conneau et al., 2019). These vectors and their combinations serve as an input to a simple feed-forward regressor which is fine-tuned to minimize the MSE loss with human MQM scores. A COMET model is trained in two stages, first on direct assessment (DA) quality annotations and then on MQM annotations, both from WMT shared tasks.

Setup. We limit our experiments to the En-De, Zh-En and Ru-En language directions because of WMT MQM availability. We largely followed the training recipe in the COMET Github repo[5]. For details, please refer to our code.

There is high inter-annotator variance in the WMT and bio MQM data. Training on the raw MQM scores is very unstable and therefore per-annotator z-normalizing is necessary to replicate our setup. Note that the publicly available WMT MQM data are not z-normalized.

[5] github.com/Unbabel/COMET/tree/master/configs

Test:WMT

DA epochs	MQM epochs 0	1	2	4	8
0	0.118	0.285	0.281	0.279	0.295
1	0.324	0.333	0.318	0.317	0.323
2	0.326	0.337	0.323	0.323	0.325
4	0.322	0.335	0.323	0.322	0.321
8	0.311	0.335	0.324	0.322	0.316

Test:Bio

DA epochs	MQM epochs 0	1	2	4	8
0	0.071	0.234	0.229	0.240	0.250
1	0.282	0.280	0.282	0.274	0.270
2	0.270	0.265	0.273	0.268	0.266
4	0.255	0.246	0.258	0.259	0.253
8	0.240	0.242	0.261	0.260	0.253

Table 3: Segment-level correlation (Kendall's τ) between metrics and human judgments on the WMT (top) and bio (bottom) test sets, for COMET with varying epochs of WMT domain DA and MQM training.
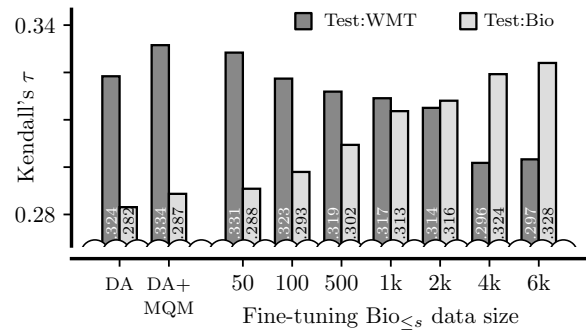

Figure 3: Average performance (8 seeds) of COMET fine-tuned on varying amounts of MQM bio data.

Observation 1: Domain gap persists throughout the fine-tuning process. We would like to understand which stage among the two training stages for COMET accounts for the domain gap. To this end, we retrained COMET with varying epochs on DA/MQM data, shown in Table 3. In contrast to catastrophic forgetting (Goodfellow et al., 2013; Thompson et al., 2019a,b), where a model starts with good general-domain performance and then overfits while being adapted to a new task or domain, we do not see a sharp dropoff in the bio domain performance when training on more WMT (DA and/or MQM) data. This indicates that the model is a weak bio metric at all stages, as opposed to first learning and then forgetting.

Observation 2: In-domain data dramatically improves COMET. Generally, including bio MQM annotations in training improves COMET's performance in the bio test set, increasing correlation
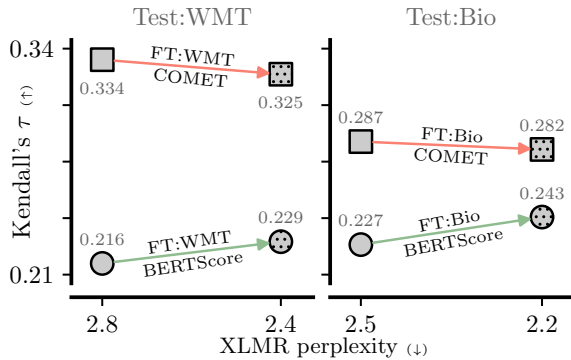
Figure 4: Metric performance when pre-trained model is fine-tuned (FT) on bio or WMT domain data. Lower perplexity improves BERTSCORE ◯ but worsens COMET ▢. Perplexity is average of MLM and TLM objectives on the text portion of the MQM dataset for both domains.

from 0.287 to 0.328 with 6k bio judgments. Indeed, just 1k judgements improves correlation to 0.313 (see Figure 3). This rules out the possibility that bio is inherently problematic for COMET's architecture or fine-tuning strategy.

4.3 How does the pre-trained model affect domain robustness?

COMET and BERTSCORE are both based on XLM-Roberta-large (Conneau et al., 2019), allowing us to explore how the same changes to the pre-trained model affect each metric. To see whether improving the underlying pre-trained model improves *Pre-trained+Algorithm* metrics built on those pre-trained models, we fine-tune XLM-Roberta with data similar to the WMT and bio domain setup, respectively. Similarly, we also investigate how PRISM, another *Pre-trained+Algorithm* metric, is affected with changes to the pre-trained model. We use PRISM with the NLLB multilingual MT models (NLLB Team et al., 2022) as they are larger and more recent than the model released with PRISM.

Setup. Our fine-tuning data covers the four languages of interest, namely English, German, Russian, and Chinese (see Appendix E.2 for a detailed data list). Since NLLB is a translation model, we use only parallel data to fine-tune the model. For the XLM-Roberta case, note that it was fine-tuned with two objectives: masked language model (MLM) and translation language model (TLM). We use both parallel and monolingual data for MLM training and parallel data for TLM training.
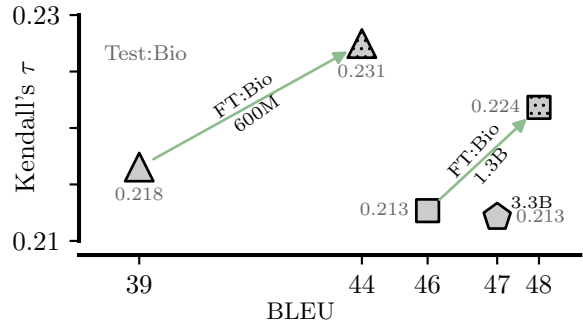

Figure 5: Multiple NLLB MT models are used as the base model for PRISMSRC. Fine-tuning the underlying MT model improves the metric. Compute constraints preclude finetuning NLLB-3.3B.

Observations: XLM-Roberta. For both domains, improving the pre-trained model improves BERTSCORE but not COMET (see Figure 4). This indicates that the limiting factor for the poor performance of COMET on bio is the effect from its various fine-tuning stages (discussed in Section 4.2), not an underlying weakness in the pre-trained model on bio.

Observations: NLLB. Our findings are shown in Figure 5. In general, we found that improving the pre-trained models performance (as measured by BLEU on a held out test set) also improved PRISM's performance.

5 Conclusion and Future Work

This paper investigated the performance of machine translation metrics across divergent domains. To this end, we introduced a new, extensive MQM-annotated dataset covering 11 language pairs in the bio domain. Our analysis showed that *Pre-trained+Fine-tuned* metrics (i.e. those that use prior human quality annotations of MT output) exhibit a larger gap between in-domain and out-of-domain performance than *Pre-trained+Algorithm* metrics (like BERTSCORE). Further experiments showed that this gap can be attributed to the DA and MQM fine-tuning stage.

Despite the gap between in-domain and out-of-domain performance, COMET is still the best performing metric on the bio domain in absolute terms. Thus, our findings suggest potential directions for future work including collecting more diverse human judgments for *Pre-trained+Fine-tuned* metrics and exploring ways to improve the generalization of such metrics during fine-tuning.

Limitations

Our findings are dependent on two empirical assumptions we discussed in section 4.1. To the best of our knowledge, those assumptions are necessary to achieve a fair comparison of metrics across domains, but conclusions may change if our assumptions are refuted in future studies.

We draw conclusions based on a single unseen domain (biomedical). While additional domains would have been preferable, data collection was cost prohibitive.

Context has been shown to be beneficial in machine translation evaluation (Läubli et al., 2018; Toral, 2020) and some metrics used in this work have document-level versions (Vernikos et al., 2022). However, in order to draw fair comparisons with existing metrics which do not yet have a document-level version, we only evaluated metrics at the sentence level.

We focused on segment-level evaluation and did not attempt system-level comparisons because of the limited number of system submissions to the WMT biomedical translation shared task.

Acknowledgements

Ethical Considerations

Our human annotations were conducted through a vendor. Annotators were compensated in accordance to the industry standard – specifically, in the range of $27.50 to $37.50 on an hourly basis, depending on the experience of the annotator.

References

Yujin Baek, Zae Myung Kim, Jihyung Moon, Hyunjoong Kim, and Eunjeong Park. 2020. PATQUEST: Papago translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 991–998, Online. Association for Computational Linguistics.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014. Machine translation quality estimation across domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 409–420, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation.

Johann Frei and Frank Kramer. 2023. German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation. *JMIR Form Res*, 7:e39077.

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

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Dam Heo, WonKee Lee, Baikjin Jung, and Jong-Hyeok Lee. 2021. Quality estimation using dual encoders with transfer learning. In *Proceedings of the Sixth Conference on Machine Translation*, pages 920–927, Online. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a

case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Aurélie Névéol, Antonio Jimeno Yepes, and Mariana Neves. 2020. MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3676–3682, Marseille, France. European Language Resources Association.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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

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

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

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

7881–7892, Online. Association for Computational Linguistics.

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.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. *arXiv preprint arXiv:2401.05749*.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019a. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019b. HABLex: Human annotated bilingual lexicons for experiments in machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1382–1387, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syn-

tactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. European Association for Machine Translation.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. 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.

| | WMT | | Bio | | |
Langs	Test	Train	Test	Dev	Total
De-En	-	-	2457	903	3360
En-De	18k	28k	2695	917	3612
Es-En	-	-	1013	309	1322
En-Es	-	-	1112	330	1442
Ru-En	-	-	1324	388	1712
En-Ru	19k	16k	825	237	1062
Fr-En	-	-	1108	352	1460
En-Fr	-	-	1228	308	1536
Zh-En	23k	27k	2838	913	3751
En-Zh	-	-	3900	1200	5100
Pt-En	-	-	701	222	924
All	60k	71k	19k	6k	25k

Table 4: Data split of the bio MQM data released in this work, and WMT22 MQM (Freitag et al., 2022) data. All test results are reported with the *test* split which is approximately 75% of *total*. Splits were created to respect document-level boundaries. For WMT, 2022 is used for testing and 2020 and 2021 for training.

A Domain Overlap Between WMT and bio

To evaluate the overlap between the WMT and bio domains, we calculate the vocabulary overlap coefficient ($\frac{|A \cap B|}{\min(|A|,|B|)}$) between our new bio MQM dataset and the domains used in the WMT22 metrics shared task. The per-domain overlap matrix is shown in Figure 6. Randomly selected sentences from each domain are provided for illustration in Figure 7.

B Corpus Statistics

Table 4 shows the size per language pair of our bio MQM dataset, as well as the WMT MQM dataset for comparison. The bio MQM dataset contains roughly 25k annotated segments, covering 11 language pairs. We split the data into test (roughly 75%) and development (roughly 25%) sets.

C Translator/Annotator Qualifications

There were 2-4 MQM annotators for each language pair, and a total of 46 annotators. All linguists had experience in translating/post-editing/reviewing content in the bio domain. This was the main requirement to be able to work on the project. The other qualification criteria for this project were in line with the ISO standard 17100. In particular, the linguists met one or more of the following criteria: (1) A recognized higher education degree in translation; (2) Equivalent third-level degree in another subject plus a minimum of two years of doc-

umented professional translation experience; (3) A minimum of five years of documented professional translation experience; (4) Native speaker of the target language. Although linguists were experts in the bio domain, not all of them were experts in MQM annotation. For this reason, the annotators completed an MQM quiz before onboarding them to ensure they understood the guidelines and requirements.

For the translation and post-editing tasks, we used a two step process (initial post editor + reviewer). In each case the reviewer was a linguist with experience translating medical texts. There were no specific educational or vocational stipulations on that medical qualification, however they were asked to provide a medical-text-specific translation test for us to be onboarded for the project. The initial post-editor in each case was a linguistic expert, but not specifically an expert in medical translations, which is why we followed up with reviewers to ensure contents were translated accurately. Linguists had to demonstrate the following to onboard to the project: (1) At least 3+ years of professional translation experience (2) Proven proficiency in English writing skills (3) In-depth understanding and exposure to the language (4) Strong ability in translating, reviewing, adjusting, and providing adaptation for various writing styles of particular requests.

D MQM Annotation Guidelines

Below, we reproduce the MQM annotation guidelines that we provided to the annotators.

Overview: You are asked to evaluate the translations using the guidelines below, and assign error categories and severities considering the context segments available.

Task:

1. Please identify all errors within each translated segment, up to a maximum of five.

    (a) If there are more than five errors, identify only the five most severe.

    (b) If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single Unintelligible error that spans the entire segment

    (c) Annotate segments in natural order, as if you were reading the document. You

	e-commerce	news	social	conversation	biomedical
e-commerce	1.000	0.349	0.511	0.662	0.369
news	0.349	1.000	0.517	0.592	0.359
social	0.511	0.517	1.000	0.494	0.462
conversation	0.662	0.592	0.494	1.000	0.554
biomedical	0.369	0.359	0.462	0.554	1.000

Figure 6: Vocabulary overlap coefficient between the English source-side data for each domain in the WMT22 and our bio dataset.

e-commerce	This was one of the first albums I purchased of Keith's "back in the day".
news	Sean Combs has been variously known as Puff Daddy, P. Diddy or Diddy, but this year announced his preference for the names Love and Brother Love.
social	The comment about boiling being inefficient is probably correct bc even though the water heater is running continuously, that thing has SO MUCH insulation.
conversation	Let me know if you were able to create your new password and sign in with it
biomedical	Though neither perfectly sensitive nor perfectly specific for trachoma, these signs have been essential tools for identifying populations that need interventions to eliminate trachoma as a public health problem.

Figure 7: Randomly selected English example sentences from each domain in the WMT22 metrics shared task as well as our new bio dataset.

may return to revise previous segments.

2. To identify an error, highlight the relevant span of text.

   (a) Omission and Source error should be tagged in the source text.

      i. All other errors should be tagged in the target text.

   (b) Unintelligible error should have an entire sentence tagged; if you think a smaller span is needed, then you should select another error category (Mistranslation, etc.).

3. Select a category/sub-category and severity level from the available options.

4. When identifying errors, please be as fine-grained as possible.

   (a) If a sentence contains more than one error of the same category, each one should be logged separately. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded.

   (b) If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe.

      i. If all have the same severity, choose the first matching category listed in the error typology (e.g. Accuracy, then Fluency, then Terminology, etc.).

   (c) For repetitive errors that appear systematically through the document: please annotate each instance with the appropriate weight.

5. Please pay particular attention to the context when annotating. You will be shown several context segments before and after the segment for evaluation. If a translation is questionable on its own but is fine in the context of the document, it should not be considered erroneous; conversely, if a translation might be acceptable in some context, but not within the current document, it should be marked as wrong.

Delivery format:

• file format: a TSV with additional columns for error categories and severity + JSON

– for multiple errors in one segment: additional row for each error + severity

– text spans will be highlighted for the annotation process and exported as tag

Error categories: Table 5

Severity (no weights, just severity): Table 6

E Supplementary Information on Experiments

E.1 Training Steps and Compute Time for Experiments

The overall training consists of the following steps (compute times using a single A10 GPU). The times are per epoch and some experiments require training for multiple epochs.

- Language modeling → XLM-Roberta, 10hr/ep.
- DA scores regression → COMETDA, 10hr/ep.
- MQM scores regression → COMET, 1hr/ep.

E.2 List of Data for Fine-Tuning Pre-Trained Model

For WMT domain, we used news-commentary v18.1 dataset[6] for all languages. For the bio domain, we list the data in Table 7.

Data Type	Language(s)	Dataset	Lines
Parallel	en-de	UFAL Medical Corpus (Yeganova et al., 2021)	3M
	en-de en-ru en-zh	MEDLINE (Yeganova et al., 2021)	35k 29k 19k
Monoling.	En	CORD (Wang et al., 2020)	1M
	De	Animal Experiments[7] GERNERMED (Frei and Kramer, 2023)	250k
	Ru	Medical QA	250k
	Zh	Chinese Medical Dataset[8]	2M

Table 7: Collection of bio domain data used in pre-trained model fine-tuning experiments.

F Raw Scores for Figure 2

The segment-level correlation (Kendall's τ) scores used to compute improvements in Figure 2 are provided in Table 8. Note that there is no public COMET 22 MQM model.

---

[6] data.statmt.org/news-commentary/v18.1/
[7] www.openagrar.de/receive/openagrar_mods_00046540?lang=en
[8] huggingface.co/datasets/shibing624/medical

Accuracy – errors occurring when the target text does not accurately correspond to the propositional content of the source text, introduced by distorting, omitting, or adding to the message	**Mistranslation**	Target content that does not accurately represent the source content.	Target
	Addition	Target content that includes content not present in the source.	Target
	Omission	Errors where content is missing from the translation that is present in the source.	*Source*
	Untranslated	Errors occurring when a text segment that was intended for translation is left untranslated in the target content.	Target
Linguistic Conventions (former Fluency) - errors related to the linguistic well-formedness of the text, including problems with, for instance, grammaticality and mechanical correctness.	**Grammar**	Error that occurs when a text string (sentence, phrase, other) in the translation violates the grammatical rules of the target language.	Target
	Punctuation	Punctuation incorrect for the locale or style.	Target
	Spelling	Error occurring when the letters in a word in an alphabetic language are not arranged in the normally specified order.	Target
	Character encoding	Error occurring when characters garbled due to incorrect application of an encoding.	Target
	Register	Errors occurring when a text uses a level of formality higher or lower than required by the specifications or by common language conventions.	Target
Terminology - errors arising when a term does not conform to normative domain or organizational terminology standards or when a term in the target text is not the correct, normative equivalent of the corresponding term in the source text.	**Inconsistent use of terminology**	Use of multiple terms for the same concept (technical terms, medical terms, etc.)	Target
	Wrong term	Use of term that it is not the term a domain expert would use or because it gives rise to a conceptual mismatch.	Target
Style	**Non-fluent**	Text does not sound fluent or natural as if it were translated by a non-native speaker or because the translation is following the source too closely.	Target
Locale Conventions - errors occurring when the translation product violates locale-specific content or formatting requirements for data elements.	**Number format**		Target
	Currency format		Target
	Measurement format		Target
	Time format		Target
	Date format		Target
	Address format		Target
	Telephone format		Target
Other		any error that does not fit the categories above	Target
Source errors	**source error**	The error that occurs in the source. All source errors (e.g. non-fluent source) should be annotated as source errors — no sub-categories need to be selected. **If the source error caused a target error:** - if the source error and target errors belong to the same category, then only flag the source. -If source and target errors belong to different categories - even if you know that the source error caused the translation error - do flag both.	*Source*
Unintelligible		So many errors, or errors are so outrageous, that text becomes incomprehensible, and it is hard to pinpoint a specific error type.	Target. Tag the entire sentence. If the span is smaller, then a different category should be applied, such as Mistranslation, Untranslated, etc.

Table 5: MQM error categories provided in annotator instructions.

severity	Definition	Source example	Translation example
Neutral	Neutral issues are items that need to be noted for further attention or fixing but which should not count against the translation. This severity level can be perceived as a flag for attention that does not impose a penalty. It should be used for "preferential errors" (i.e, items that are not wrong, per se, but where the reviewer or requester would like to see a different solution).	Source: Join us in celebrating 10 years of the company!	Target: Join us to celebrate 10 years of the company!
Minor	Minor issues are issues that do not impact usability or understandability of the content. If the typical reader/user is able to correct the error reliably and it does not impact the usability of the content, it should be classified as minor.	S1: Accurately distinguish between legitimate and high-**r**isk account registrations S2: See how organizations worldwide are using fraud detection.	T1: Accurately distinguish between legitimate and high- **r**isk account registrations T2: See how organization worldwide are using fraud detection.
Major	errors that would impact usability or understandability of the content but which would not render it unusable. For example, a misspelled word that may require extra effort for the reader to understand the intended meaning but does not make it impossible to comprehend should be labeled as a major error. Additionally, if an error cannot be reliably corrected by the reader/user (e.g., the intended meaning is not clear) but it does not render the content unfit for purpose, it should be categorized as major.	Source: Set the performance to 50 percent	Target: Set performance 50 percent
Critical	errors that would render a text unusable, which is determined by considering the intended audience and specified purpose. For example, a particularly bad grammar error that changes the meaning of the text would be considered Critical. Critical errors could result in damage to people, equipment, or an organization's reputation if not corrected before use. If the error causes the text to become unintelligible, it would be considered Critical.	S1: Set the device on the highest temperature setting. S2: The next step would be to identify the point of leakage. S3: 1.3 degrees	T1: Set the device on the lowest temperature setting. T2: It would be to identify the next point of leakage. T3: 1,300 degrees

Table 6: Severity examples and explanations provided in MQM annotation instructions.

Type	Metric	Test:WMT	Test:Bio
Surface-Form	BLEU	0.134	0.213
	ChrF	0.151	0.192
	TER	0.140	0.100
Pre-trained+Algorithm	PRISM$_{REF}$	0.216	0.242
	PRISM$_{SRC}$	0.121	0.267
	BERTScore	0.216	0.227
Pre-trained+Prompt	GEMBA$_{DAV3}$	0.280	0.159
	GEMBA$_{DAV3.QE}$	0.222	0.173
Pre-trained+Fine-tuned	COMET$_{MQM.21}$	0.328	0.249
	COMET$_{QE.21}$	0.294	0.205
	COMET$_{DA.21}$	0.309	0.284
	COMET$_{INHO.21}$	0.255	0.182
	COMET$_{DA.22}$	0.304	0.269
	UniTE	0.301	0.249
	BLEURT	0.214	0.100

Table 8: Segment-level correlation (Kendall's τ) between metrics and human judgments on the WMT and bio domain. *Pre-trained+Fine-tuned* metrics have lower correlation on bio than on WMT, while *Surface-Form* and *Pre-trained+Algorithm* tend to have higher correlation.