Assessing the Role of Context in Chat Translation Evaluation: Is Context Helpful and Under What Conditions?

Sweta Agrawal[1*], **Amin Farajian**[2], **Patrick Fernandes**[1,5,6]
Ricardo Rei[1,2,3], **André F. T. Martins**[1,2,4,5]
[1]Instituto de Telecomunicações, Portugal [2]Unbabel, Portugal [3]INESC-ID, Portugal
[4]ELLIS Unit Lisbon, Portugal [5]Instituto Superior Técnico & Universidade de Lisboa, Portugal
[6]Carnegie Mellon University, USA
swetaagrawal20@gmail.com

Abstract

Despite the recent success of automatic metrics for assessing translation quality, their application in evaluating the quality of machine-translated chats has been limited. Unlike more structured texts like news, chat conversations are often unstructured, short, and heavily reliant on contextual information. This poses questions about the reliability of existing sentence-level metrics in this domain as well as the role of context in assessing the translation quality. Motivated by this, we conduct a meta-evaluation of existing automatic metrics, primarily designed for structured domains such as news, to assess the quality of machine-translated chats. We find that reference-free metrics lag behind reference-based ones, especially when evaluating translation quality in out-of-English settings. We then investigate how incorporating conversational contextual information in these metrics for sentence-level evaluation affects their performance. Our findings show that augmenting neural learned metrics with contextual information helps improve correlation with human judgments in the reference-free scenario and when evaluating translations in out-of-English settings. Finally, we propose a new evaluation metric, CONTEXT-MQM, that utilizes bilingual context with a large language model (LLM) and further validate that adding context helps even for LLM-based evaluation metrics.

1 Introduction

Automatically estimating the quality of machine or human-generated translations has received a lot of attention over the past two decades from the NLP community (Han et al., 2021), specifically via shared tasks organized by WMT from 2014–present (Macháček and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016, 2017; Ma et al., 2018, 2019; Mathur et al., 2020; Freitag et al., 2021, 2022, 2023). A variety of evaluation metrics have been developed for this purpose, encompassing lexical matching approaches such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015); embedding-based methods like BERTSCORE (Zhang et al., 2019) and Word Mover Distance (Zhao et al., 2019); learned metrics like COMET (Rei et al., 2020a), and BLEURT (Sellam et al., 2020); and metrics that employ prompting techniques with large language models (LLMs) like GEMBA-MQM (Kocmi and Federmann, 2023a) or AUTOMQM (Fernandes et al., 2023a).

Among these metrics, neural metrics have gained widespread acceptance (Freitag et al., 2022) as they are directly trained to predict sentence-level translation quality assessment scores (Kreutzer et al., 2015; Rei et al., 2020a; Sellam et al., 2020), word-level error annotations collected by professional linguists (Guerreiro et al., 2023), or post-editing efforts as measured by HTER (Snover et al., 2006; Fonseca et al., 2019; Specia et al., 2021). However, the reliance on human-written reference translations and judgments collected predominantly from structured domains like news or Wikipedia as training data raises questions about their adaptability and reliability in detecting errors in other domains (Zouhar et al., 2024), for example, in evaluating translation quality in more informal settings.

Unlike news articles, which involve carefully authored and well-formatted text, and which current translation systems are well equipped for, chat conversations are often synchronous and short, and involve formal language, colloquial expressions, and slang that may not have direct equivalents in the target language (Gonçalves

[*]Work partially developed during internship at Unbabel.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1250–1267, 2024. https://doi.org/10.1162/tacl_a_00700
Action Editor: David Chiang. Submission batch: 3/2024; Revision batch: 6/2024; Published 9/2024.

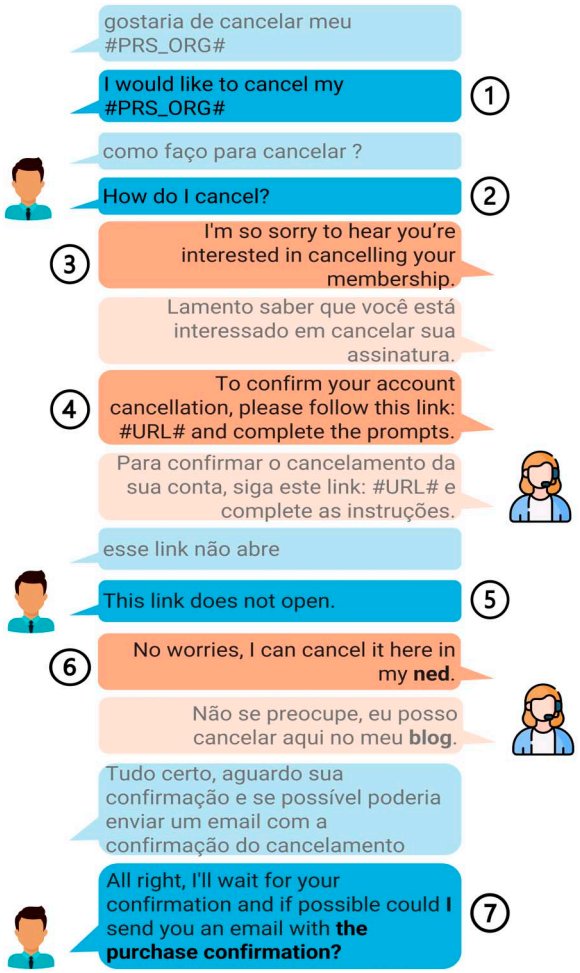


Figure 1: An example chat from the MAIA corpus (Martins et al., 2020): The agent and the customer only see the texts in their respective languages. Errors both MT and user-generated are bold-faced.

et al., 2022). An example of such a conversation is presented in Figure 1. Although chat messages are relatively easier to translate, errors introduced by machine translation (MT) systems might go unnoticed by end users if not detected properly, potentially leading to miscommunication, conversation breakdowns, or even more serious implications on high-risk domains (e.g., patient-physician chats) (Yamashita et al., 2009; Robertson and Díaz, 2022). Identifying these errors *as they occur* can potentially help the users ask clarification questions and correct any misunderstandings due to the mistranslation of the intended information (Gao et al., 2015).

Moreover, conversational texts rely heavily on context, meaning that the interpretation of a text is largely influenced by the surrounding contextual information. Hence, MT systems for such domains are often trained with contextual information and this has been shown to improve translation quality, lexical inconsistency, and coherence of the generated outputs (Farinha et al., 2022; Fernandes et al., 2021). However, it is unclear to what extent context plays a role in estimating translation quality for machine-translated conversations.

In this work, we first systematically analyze the nature and the frequency of the errors in real bilingual chat translations from customer support and contrast them with the structured news domain (§ 2). We find that translation errors are 21% less frequent in chat relative to the news domain and that the nature of the errors introduced by MT systems in the two domains is also different. This underscores the importance of understanding and evaluating existing automatic metrics for chat translation. Due to the infrequent error occurrences at the turn level, MT systems for chat translation tasks might receive higher scores, necessitating a more nuanced evaluation of the metrics.

Motivated by the same, we present a meta-evaluation of automatic metrics, primarily tested on news translation tasks, in their ability to gauge the quality of machine-translated chats at turn and dialogue level (§ 3). We use the Multidimensional Quality Metrics (MQM) annotations collected by the WMT22 Chat Shared Task on systems submitted for bilingual customer chat translations (Farinha et al., 2022), where experts were asked to rate the quality of translation *at each turn* given the bilingual context, independently for all language directions. We evaluate the sentence-level metrics across two scenarios: on *all* translation pairs as well as *imperfect* translations as judged by humans. We then study the impact of augmenting a subset of the learned sentence-level neural metrics with different types of conversational contextual information (§ 4). Our findings are summarized below:[1]

- COMET-22, a reference-based metric, best correlates with human judgments.
- Reference-based COMET-22 does not benefit from the added contextual information, whereas reference-free COMET-20-QE has better correlations with human judgments as

[1]Code is available at https://github.com/sweta20/chat-qe.

the context window increases when evaluating translations in out-of-English settings. This is useful as reference-free metrics allow translation quality to be assessed on the fly.

- Adding context helps assess translation quality better for short and ambiguous sentences. Using correct and complete context from both speakers improves chat quality estimation via COMET-based metrics.

Finally, we present CONTEXT-MQM, an LLM-based evaluation metric that uses context for chat translation quality estimation (§ 5.3). Our preliminary experiments with CONTEXT-MQM show that adding bilingual context to the evaluation prompt helps improve the quality assessment of machine-translated chats on imperfect translations.

2 Errors in Chat vs. News: A Case Study

To better understand how the error types differ in these domains, we present an analysis of the nature and the frequency of errors using MQM annotations (conversation: 7120, news: 4800 translation pairs) collected for English-German as a part of the WMT22 Metrics shared task (Freitag et al., 2022).

Errors are Less Frequent. We calculate the percentage of translations with a perfect MQM score from the news and conversational subset of the WMT22 dataset: only 46.4% of news translations have perfect MQM scores, whereas this percentage is 57.8% for the conversational domain. This suggests that errors are less frequent in the conversational domain, likely due to the relatively short (see Figure 2) and probably less complex text dominant in conversations. However, it is important to note that these errors do not occur in isolation and can escalate into larger communication issues. Moreover, if quality estimation metrics fail to accurately detect less frequent errors, they may misrepresent the true quality of the MT systems.

Most Frequent Error Types Across Domains are Different. Errors related to fluency, such as issues with spelling, consistency, and register, occur more frequently in conversations compared to accuracy-based errors like mistranslation, which

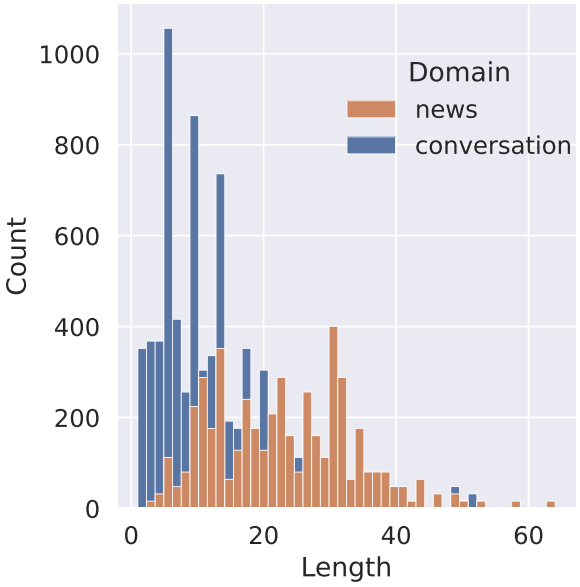


Figure 2: Conversational texts tend to be much shorter than news texts in WMT22 EN-DE.

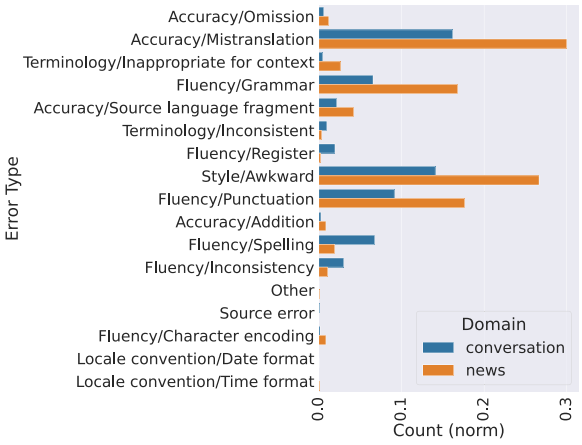


Figure 3: Counts of MQM error categories normalized by the number of annotated instances for each domain: frequent errors differ in the two domains.

are more common in the news domain (Figure 3). This underscores the idea that the nature of observed errors is also influenced by the specific domain context.

The variation in error types and their frequencies across the two domains underscores the need for a systematic study of the suitability of current automatic metrics within the chat domain.

3 Meta Evaluation of Automatic Metrics

We now investigate to what extent automatic MT metrics capture chat translation quality. We

LP	Sender	Count			Length			CUA				% Perfect MQM	# Noisy Source[3]
		# (Instances)	# (Chats)	# (Systems)	Source	MT	Ref	Weak	Moderate	Good	Excellent		
En-De	A	2715	30	5	48.82	57.34	57.20	89	90	120	2219	73.6	55
De-En	C	2720			38.14	34.46	35.49	113	88	166	2067	65.4	105
En-Fr	A	1868	23	2	36.26	42.81	44.21	100	118	129	1274	46.7	40
Fr-En	C	1020			35.72	32.36	33.19	45	37	41	748	65.7	116
En-Pt	A	1318	28	2	43.21	46.34	45.96	77	81	103	872	50.1	24
Pt-En	C	1016			30.44	31.06	31.35	40	68	66	624	54.5	95

Table 1: Statistics from the MQM annotations of the WMT 22 Chat Shared Task: Errors are less frequent in the chat domain with 46.7% to 73.6% translations receiving perfect MQM scores. A: Agent; C: Customer.

detail the setup used for our meta-evaluation below:

3.1 Dataset

We use the MQM annotations collected from the WMT 2022 Chat Shared Task.[2] The dataset consists of genuine bilingual customer support conversations translated by participants' submitted MT systems. The translations were evaluated at turn-level using the whole conversational context via an MQM-based human evaluation framework.

We convert token-level MQM spans into a turn-level score via the following formula:

$$\text{MQM} = -(C_{\text{Min}} + 5 \times C_{\text{Maj}} + 10 \times C_{\text{Cri}}) \quad (1)$$

where, C_{Min}, C_{Maj}, and C_{Cri} denote the number of minor, major, and critical errors, respectively (Lommel et al., 2014; Farinha et al., 2022). We measure the dialogue-level translation quality as the mean turn-level MQM scores.[3]

Table 1 shows several statistics extracted from the data.[4] Across language pairs, the percentage of instances with no errors (% Perfect MQM) ranges from 46.7 to 73.6%, confirming our initial analysis that errors are less frequent in translated chats. Following Farinha et al. (2022), we also present Customer Utility Analysis (CUA) bucketing MQM scores in four regions: Weak (negative - 39); Moderate (40 – 59); Good (60 – 79) and Excellent (80 – 100). It is apparent that translating agent directions results in higher quality translations than translating customer directions, potentially due to less noisy source texts (Gonçalves et al., 2022).

3.2 Automatic Metrics

We benchmark sentence-level and document-level metrics frequently used for translation quality assessment. Following the WMT QE shared task evaluation (Blain et al., 2023), we report Spearman-rank correlation (Zar, 2005) to measure how well these metric scores align with human judgments.[5]

3.2.1 Sentence-level

BLEU (Papineni et al., 2002) estimates the translation quality based on n-gram overlap between the hypotheses and references. We compute sentence-level BLEU (Chen and Cherry, 2014) using the SacreBLEU (Post, 2018) library.[6]

chrF (Popović, 2015) evaluates the similarity by computing an F1 score between the overlapping character n-grams in the hypotheses and references.

BERTScore (Zhang et al., 2019) computes the cosine similarity between the pre-trained contextualized embeddings of hypotheses and references.

BLEURT (Sellam et al., 2020) uses pre-trained transformer models to estimate the semantic similarity between the hypothesis and the reference. Its is a neural regression metric trained on an existing collection of human judgments.

Comet-22 (Rei et al., 2022a) is an XLM-R-based (Conneau et al., 2020) regression metric trained

[2] https://github.com/WMT-Chat-task/data-and-baselines.

[3] While simple averaging of turn-level scores does provide a dialogue-level baseline, it might be insufficient for capturing the true quality of translations at the dialogue level that goes beyond turn-level errors.

[4] Note that the Agent communicates always in English and the Customer in non-English languages.

[5] In this work, we opted for Spearman since it serves as a compromise between Pearson and Kendall (Deutsch et al., 2023).

[6] https://github.com/mjpost/sacrebleu/: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0.

on direct assessments from WMT17 to WMT20. Unlike BLEURT which compares the hypothesis against the reference, COMET returns a score, measuring the quality of translsation relative to both source and reference.

COMET-20-QE (Rei et al., 2020b) is also an XLM-R-based regression metric trained to predict the direct assessment scores using representations extracted from the (source, hypothesis) pair.

COMETKIWI-22-QE (Rei et al., 2022b) is a reference-free model built on top of the INFOXLM-R model (Chi et al., 2021) and is trained to predict direct assessments from WMT17-20 and the MLQE-PE corpus (Fomicheva et al., 2022).

XCOMET-XL (Guerreiro et al., 2023), with 3.5B parameters, is an explainable learned metric trained to predict both sentence-level quality scores and MQM-like error spans from the (source, hypothesis, reference) triplet. It can also be used for quality estimation by using only the source and the hypothesis as input, referred to as XCOMET-QE-XL.

METRICX-23-XL (Juraska et al., 2023) is a regression-based metric built on top of mT5 (Xue et al., 2021) and is trained to regress the true MQM score to predict an error score in the range [0, 25]. The input to the model is the concatenated hypothesis and reference translations.

METRICX-23-QE-XL (Juraska et al., 2023) is a reference-free version of METRICX-23-XL that instead takes as input the source and the hypothesis.

3.2.2 Document-level

Following Liu et al. (2020), we compute dialogue-level BLEU (**d-BLEU**) considering n-grams over all the turns in a dialogue when comparing against the reference n-grams. We compute the average of turn-level **COMET-22**, **COMET-20-QE**, and **COMETKIWI-22-QE** scores within a dialogue to represent its translation quality. Finally, we also evaluate **SLIDE** (Raunak et al., 2023), a document-level reference-free metric that computes translation quality by aggregating metric scores over a block of sentences. Following Raunak et al. (2023), we use COMETKIWI-22-QE on a moving window of block size $k = 6$ with a stride of 6.

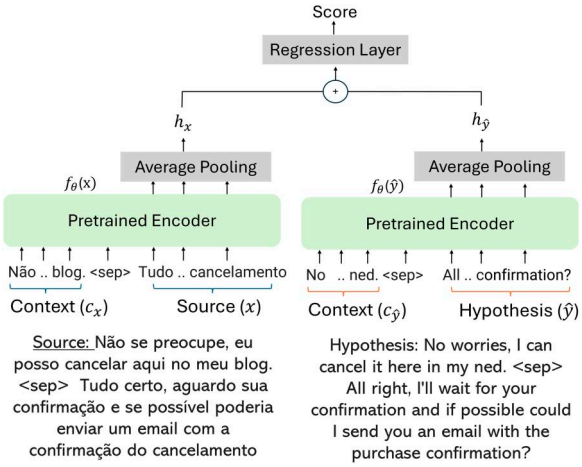


Figure 4: CONTEXT COMET-QE.

4 Context-Aware Translation Evaluation

Let Agent (A) and Customer (C) represent the two participants in a bilingual chat. Given a text x generated by A or C, the goal is to predict the quality of its translation, $\hat{y}$, given an optional reference translation, y. We extend COMET-based metrics to utilize conversational context as detailed below:

Incorporating Context COMET uses pooled token-level representations extracted from pre-trained language models (parameterized by f_θ) like BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) to encode source sentences (x), hypotheses ($\hat{y}$) and the reference translations (y) into sentence-level contextual vectors ($[f_\theta(x), f_\theta(\hat{y}), f_\theta(y)]$). The concatenated feature vectors are then passed through a regression layer to generate a quality score. Following Vernikos et al. (2022), we obtain the token-level contextual representations of the **current** source, x, and the reference (y)/hypothesis ($\hat{y}$) sentences by prepending up to k sentences of context preceding to it. However, only the representations of the current instance are pooled before passing them to the regressor module that generates the quality score. We illustrate this for the reference-free COMET-20-QE that uses source and hypothesis to produce a quality score in Figure 4 for a context window of size 1. For COMET-22, the regression layer can also access the reference vector (h_y).

This simple approach was used to evaluate document-level translation quality and was shown to be more effective over sentence-level counterparts (Vernikos et al., 2022), whereas, in our

	Metric	Average		En-De		En-Fr		En-Pt	
		All	Imperfect	All	Imperfect	All	Imperfect	All	Imperfect
Ref-based	chrF	0.531	0.404	0.430	0.253	0.666	0.510	0.496	0.448
	BLEU	0.500	0.319	0.363	0.151	0.642	0.426	0.494	0.381
	BERTScore	0.545	0.403	0.439	0.250	0.680	0.501	0.516	0.457
	BLEURT	0.372	0.511	0.398	0.411	0.326	0.663	0.393	0.458
	Comet-22	**0.633**	**0.551**	**0.578**	**0.445**	0.721	0.634	0.602	**0.573**
	xComet-XL	0.402	0.525	0.402	0.432	0.381	0.649	0.422	0.494
	MetricX-23-XL	0.622	0.543	0.535	0.404	**0.723**	**0.692**	**0.609**	0.535
Ref-Free	Comet-20-QE	0.379	0.316	0.410	0.294	0.376	0.339	0.351	0.315
	CometKiwi-22-QE	0.300	**0.416**	0.366	0.416	0.203	**0.508**	0.331	0.325
	xComet-QE-XL	0.276	0.363	0.349	0.376	0.160	0.383	0.318	0.329
	MetricX-23-QE-XL	**0.447**	0.402	**0.438**	**0.371**	**0.437**	0.479	**0.466**	**0.356**

Table 2: Comet-22 achieves the highest correlation on average across all **Agent** language pairs with MetricX-23-XL as a close competitor in the Ref-Based setup. MetricX-23-QE-XL outperforms Comet alternatives in the ref-Free setting.

work, we study and extend its applicability to contextualized sentence-level chat quality estimation. Furthermore, this approach can be utilized with any metrics that aggregate or use contextualized token-level representations to generate a sentence-level representation, *e.g.*, BertScore.

Choice of Context We explore the usage of two types of contextual information for translation quality estimation. The current text, assuming it is generated by Customer C, can be preceded by the context from the same participant or the context generated by the other participant. For example, for the source text generated by Customer C at time $t = 7$ in Figure 1, the two preceding contextual sentences ($k = 2$) are shown:

> Context:
> *Customer at* $t = 5$
> Original: esse link não abre
> Translation: This link does not open.
> *Agent at* $t = 6$
> Original: No worries, I can cancel it here in my ned.
> Translation: Não se preocupe, eu posso cancelar aqui no meu blog.

We prepend up to k sentences of source and translation context to the current source (x) and the hypothesis ($\hat{y}$)/reference (y) separated by a tag, <sep> in the *within* setting. For the *across* setting, to have the context in the same language on each side, we prepend the translated context to the source (x) and the source context to the hypothesis ($\hat{y}$)/reference (y) from the *other* participant. Using our proposed extensions of the sentence-level metrics (Comet-22 and Comet-20-QE), we study their impact on translation quality evaluation.

5 Results

We first present the results of the meta-evaluation of existing automatic MT metrics. Then we discuss the impact of adding context to a subset of sentence-level metrics with ablations on how context impacts translation quality evaluation. Finally, we present a preliminary study on utilizing context with LLM-based MT evaluation.

5.1 Meta Evaluation

Tables 2 and 3 show the correlation of human judgments at the turn-level with automatic metrics for all the Agent and the Customer language pairs respectively. ''Imperfect'' translations are instances marked with an MQM score of <0. Table 4 reports the correlation of dialogue-level metrics with conversation-level translation quality assessment.

Turn-level Evaluation When considering all the instances in the corpus (''All''), Comet-22 achieves the highest correlation on both settings (Agent and Customer), outperforming all other metrics, with MetricX-23-XL as a close competitor. For reference-free evaluation, MetricX-23-QE-XL and Comet-20-QE achieve the highest

	Metric	Average		De-En		Fr-En		Pt-En	
		All	Imperfect	All	Imperfect	All	Imperfect	All	Imperfect
Ref-Based	chrF	0.427	0.188	0.400	0.201	0.411	0.157	0.469	0.205
	BLEU	0.396	0.166	0.390	0.154	0.373	0.147	0.425	0.198
	BERTScore	0.484	0.280	0.445	0.239	0.467	0.332	0.539	0.269
	BLEURT	0.559	0.451	0.540	0.445	0.520	0.464	0.617	0.444
	Comet-22	**0.610**	0.517	0.580	0.438	**0.588**	0.535	**0.661**	0.578
	xCOMET-XL	0.454	**0.566**	0.357	**0.514**	0.479	**0.588**	0.527	0.594
	MetricX-23-XL	0.608	0.551	**0.589**	0.511	0.583	0.544	0.651	**0.598**
Ref-Free	Comet-20-QE	**0.516**	0.415	**0.554**	0.381	0.471	0.429	0.523	0.435
	CometKiwi-22-QE	0.438	0.443	0.385	0.463	0.456	0.406	0.473	0.461
	xCOMET-QE-XL	0.447	0.493	**0.388**	**0.492**	0.462	**0.479**	**0.492**	0.508
	MetricX-23-QE-XL	0.395	**0.497**	0.383	0.490	0.382	0.431	0.420	**0.569**

Table 3: On average, on the **Customer** language pairs, Comet-22 and MetricX-23-XL achieve the highest correlation scores and neural learned metrics consistently outperform lexical metrics.

Metric	En⇔De	En⇔Fr	En⇔Pt-Br
dial-BLEU	0.657	0.845	0.708
Comet-22	**0.705**	**0.936**	**0.844**
Comet-20-QE	0.583	0.583	0.462
CometKiwi-22-QE	0.568	0.819	0.647
Slide (6, 6)	**0.611**	**0.823**	**0.656**

Table 4: Dialogue-level Evaluation of Metrics.

correlation with human judgments when evaluating translations out of English and into English respectively. However, there is a big gap between the best-performing metric for the reference-free and reference-based evaluation on average across all Agent (δ(Comet-22,MetricX-23-QE-XL): 0.186) and Customer (δ(Comet-22,Comet-20-QE): 0.094) language pairs.

There is no clear winner for imperfect translations: most metrics suffer a drop in correlation for this subset compared to ''All'' translations except xComet-XL and xComet-QE-XL. xComet-XL consistently achieves better correlations on this subset than ''ALL'' data. This could be due to an over-prediction of errors via the metric for chats.

Neural reference-based metrics consistently outperform lexical metrics in most settings, specifically when evaluating translations into English. However, when assessing the translation quality in out-of-English language pairs (Agent), reference-based lexical metrics achieve better correlations with human judgments than reference-free metrics, suggesting room for improvement for reference-free evaluation for assessing translations in languages other than English.

Dialogue-level Evaluation Comet-22 outperforms lexical metric d-BLEU in evaluating translations at the conversation-level. Interestingly, Slide (6, 6) achieves close correlation scores to Comet-22 showing its efficacy in evaluating dialogues when references are unavailable.

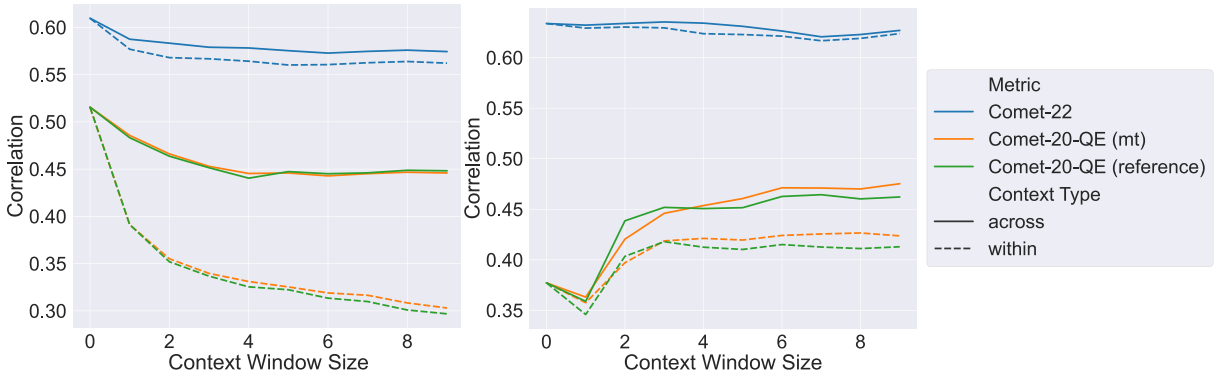
5.2 Context-Aware Translation Evaluation

We consider the context-aware extensions of reference-based Comet-22 and reference-free Comet-20-QE. For each of these metrics, we study the impact of adding contextual information as detailed in § 4 in both *within* or *across* participants settings. For reference-free Comet-20-QE, we additionally consider the setup where we use the machine-translated hypothesis instead of the reference as context. This is to mimic the real-word chat scenario where references are generally unavailable. We hypothesize that noisy context can still provide useful information in estimating the quality of the current (source, translation) pair.

5.2.1 Main Results

Figure 5 shows results of adding up to last nine sentences as context to the above configurations averaged across customer (''Average Customer'') and agent (''Average Agent'') language pairs.

Context is not helpful when references are available. Reference-based Comet-22 on average across all language pairs does not benefit (Agent) or hurts (Customer) correlation with the added context information. This could be because most of the necessary information to resolve any



(a) Into-English (Customer) (b) Out-of-English (Agent)

Figure 5: Impact of varying context window and context type (across/within) on average correlation across Agent and Customer settings: adding context helps improve metrics performance in out-of-English reference-free settings (Agent) but is detrimental for into-English (Customer) evaluation.

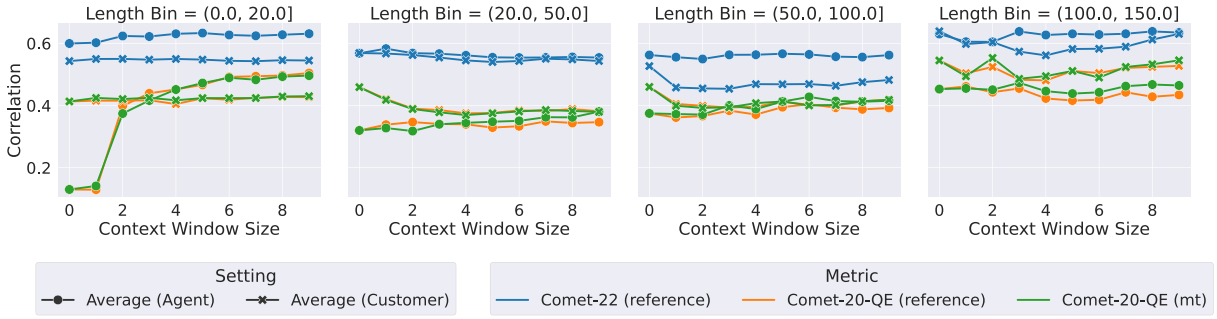


Figure 6: Context helps the most in improving the assessment of shorter (source length $\leq$ 20) and potentially ambiguous ‘‘Agent’’ (English) source sentences in *across* setting.

ambiguity is ideally already included in the reference. Adding more context could potentially introduce ambiguity or inconsistency that is not present in the reference text, hurting the evaluation process.

Adding context negatively impacts the evaluation of translations into English. For all customer directions, adding context almost always hurts the metric's correlation with human judgments. Even in the reference-free scenario, translation evaluation into English does not benefit from the added context. We hypothesize that this could be due to how contextual ambiguities are expressed in English compared to other languages and how the QE metric handles context in these languages. We leave further investigation to future work.

Adding context improves correlation for COMET-20-QE in reference-free out-of-English settings. Reference-free COMET-20-QE significantly benefits from the added context on average across all ‘‘Agent’’ settings. The correlation increases as the context increases. Specifically, using complete contextual information from both participants in the same language as the current participant (*across*) is key to getting the most out of the added contextual information. Shorter segments ($\leq$ 20 characters) benefit the most from the added context as depicted in Figure 6. The above trend holds when using either reference-based or hypothesis-based contexts, which is promising.

5.2.2 Ablation Analysis

We use the *across* participant setting with a context window of 2 for COMET-20-QE, as this setup led to the most improvement in correlation. In addition, we use the hypothesis as the context instead of the reference for all the analysis, mimicking the real-world scenario where references are unavailable. With this setup, we first study whether adding context improves correlation for

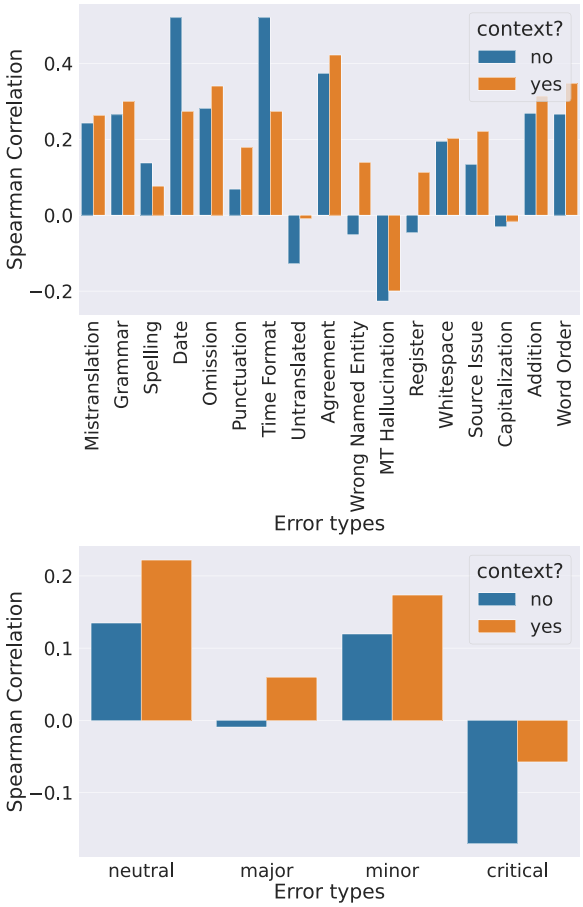


Figure 7: Adding context improves correlation with human judgments across most error types (except Date and Time Format) and all severity levels.

specific error types and severity levels. We then evaluate the impact of adding noise to the context with the proposition that unrelated or partial context should hurt the metric's performance.

Breakdown by Error Types and Severity We bucket sentences by the MQM error typology and filter error types with at least 20 instances. Figure 7 shows the correlation with human judgments for each error type bucket and across severity levels. Including contextual information improves correlation with human judgments on several error types as well as all error severity levels. However, adding context hurts Spelling, Date, and Time Format errors, possibly because adding context might introduce noise and not provide any useful information to identify such errors which are solely based on linguistic or formatting rules.

Impact of Noisy or Partial Context To validate that complete and correct context is necessary for meaningful improvement in metrics' performance, we inject two types of noise into the context: **Swap**, where we use unrelated context from a different instance; and **Drop** where we drop one of the two (source, translation) pairs (''pair'') or unpaired sentences from the preceding context (''random''). We additionally consider injecting these noises into either the source, the translation, or both.

	Noise to?	Avg-Agent
No Context	–	0.379
Context	–	0.420
Swap	Source	0.364
	Translation	0.296
	Both	0.299
Drop (Random)	Source	0.402
	Translation	0.326
	Both	0.346
Drop (Pair)	Source	0.389
	Translation	0.399
	Both	0.325

Table 5: Corrupting context hurts correlation.

Table 5 shows that adding either kind of noise to the context leads to a drop in correlation relative to the ''Context'' baseline, often even performing worse than using any contextual information (''No Context''). Unrelated context (''Swap'') has a more adverse impact on the metric's performance compared to changing the context via dropping partial information. Furthermore, dropping paired contextual sentences results in a larger drop in correlation than dropping unrelated source-translation instances. This further solidifies our argument that complete and related context is key to utilizing context for chat translation quality estimation.

Impact on Contextually Ambiguous Sentences Given that shorter sentences benefit the most from the added context (Figure 6), we further investigate whether this is indeed due to the increased ambiguity in texts. We use MuDa (Fernandes et al., 2023b) to identify translation pairs with specific discourse phenomena (formality, pronouns, verb form, lexical consistency) for English-German. This enables us to mark instances that potentially require context for disambiguation. Figure 8 shows that the sentences

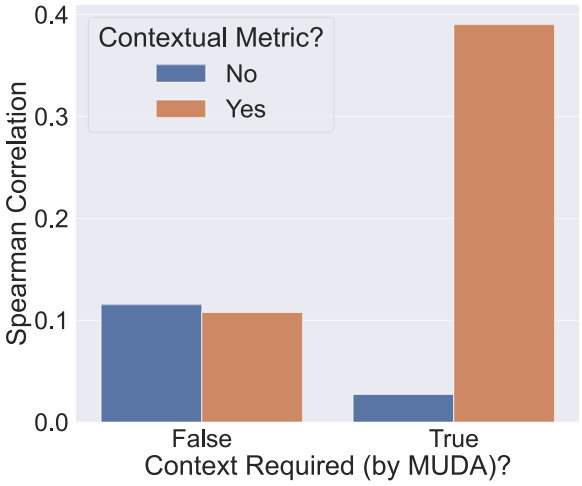


Figure 8: Adding context improves correlation on contextually ambiguous sentences (character length ≤ 20) for English-German.

with marked discourse phenomena benefit the most from the added context. On the other hand, adding context to sentences without marked discourse phenomena hurts correlation. We leave the detailed exploration of when to rely on context for quality estimation to future work.

5.3 LLM-based Contextual Quality Estimation

Motivated by the recent surge of interest in using LLMs like GPT (Achiam et al., 2023) for text evaluation, specifically in MT (Kocmi and Federmann, 2023b; Fernandes et al., 2023a; Lu et al., 2023; Kocmi and Federmann, 2023a), we also explore their potential in assessing the translation quality of machine translated chats. To better elicit the reasoning and in-context learning capabilities of these LLMs, we prompt GPT-4[7] to identify and categorize errors in machine-generated translations instead of asking for an overall score for translation quality like direct assessment.[8]

We present CONTEXT-MQM, a context-informed LLM-based quality estimation metric for chat translation evaluation. We adopt the MQM-style prompting techniques to elicit the reasoning capabilities of the LLM (Fernandes et al., 2023a; Kocmi and Federmann, 2023a) and modify it to utilize contextual information: Following Section 5.2, we include the past $k = 8$ bilingual source sentences as context and one in-domain in-context example as shown in Figure 9. We perform our evaluation on a subset of 1000 English-German sentences sampled uniformly from the dataset and contrast our metric with the evaluation prompting technique that does not utilize any contextual information, LLM-MQM (NO CONTEXT).[9]

The results are presented in Table 6: Adding context positively impacts correlation for LLM-based evaluation of machine-translated chats. CONTEXT-MQM improves correlation with human judgments, outperforming both non-contextual LLM-MQM (All: +0.013, Imperfect: 0.048) as well as COMET-22 (All: +0.091, Imperfect: +0.107). The improvement is larger on the imperfect translations, suggesting that context helps identify errors better on these (source, translation) pairs. These initial results show the potential of using LLMs for evaluating chat translation quality with contextual information. Exploring alternative prompting strategies to integrate context in LLMs merits further investigation in future research endeavors.

6 Related Work

Automatic MT Metrics Designing automatic metrics to assess translation quality has been an active area of research over the past decade. Metrics shared tasks organized at WMT have significantly facilitated research where recent metrics like BLEURT (Sellam et al., 2020) or COMET (Rei et al., 2020b) based on neural architectures and trained with human assessments are shown to consistently outperform lexical metrics. Recent work has also focused on developing document-level evaluation metrics acknowledging that sentences often do not occur in isolation in the wild and the correctness of translation is dependent on the context (Voita et al., 2019). Document-level metrics like SliDe (Raunak et al., 2023) or BlonDe (Jiang et al., 2022) use discourse information to assess the translation quality at the paragraph level. However, these metrics haven been primarily evaluated for assessing the quality of news-like data.

Chat Translation Quality Estimation Li et al. (2022) introduce the erroneous chat translation

[7] `gpt-4-0613`, accessed on February 26, 2024.

[8] Our initial experiments with open-sourced LLMs suggested limited ability of the models to do well on this task.

[9] We did not conduct a full-scale evaluation due to the high cost of accessing the GPT-4 API. The cost for running the experiments on English-German was $\sim$ $300.

System: You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the translation.

The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, source error or no-error.

Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

{k Few Shot Incontext Examples}

User: Context: "'{context}"'

{sender} source (source_lang): "'{source_seg}"'

{target_lang} translation: "'{target_seg}"'

Based on the conversation context between the agent and the customer, the current source by "sender" in {source_lang} and its machine translation in {target_lang} surrounded with triple backticks, identify error types in the translation and classify them.

Figure 9: Contextual Prompt for Chat Quality Estimation.

	All	Imperfect
LLM-MQM (No Context)	0.642	0.512
Context-MQM	0.655	0.560
Comet-22	0.564	0.453

Table 6: Context-MQM outperforms Comet-22 and non-contextual LLM-MQM on English-German Chat Quality Estimation.

detection task and propose an error detection model that classifies a given translation in a bilingual *two-utterance* chat as either correct or erroneous. However, their approach requires training a metric on chat data, whereas, in our work we benchmark existing metrics for chat quality estimation and study the impact of conversation context on quality estimation via existing metrics.

Menezes et al. (2023) propose a new framework for identifying contextual errors in conversational datasets. They expand the MQM categories to account for errors introduced due to *contextual triggers*. They further show that these errors are indeed critical and that current metrics fall short in detecting them. Our evaluation instead targets estimating chat translation quality regardless of the specific errors and is aimed at providing a more general view of the ability of existing metrics to assess the quality of machine-translated chats.

Contextual Machine Translation In many scenarios, translation requires leveraging information beyond the sentence level to resolve inter-sentence dependencies and improve translation quality. Incorporating context to generate high-quality translations has been explored for conversation and news documents, with approaches ranging from simply concatenating the context to the original input (Tiedemann and Scherrer, 2017) to more complex options (Jean et al., 2017; Maruf et al., 2018, 2019). However, despite having access to context, contextual MT models still struggle to effectively use it (Fernandes et al., 2021), and most MT metrics fail to capture this due to a lack of utilization of context in the evaluation metrics themselves. Hence, we instead use context to assess the translation quality of a given (source, target) pair and show that it benefits the evaluation of non-English translations.

7 Conclusion and Discussion

Estimating translation quality in diverse domains is crucial to ensure that the metrics employed in MT evaluation accurately reflect the MT system's quality across various types of content. We show that the nature and the type of errors in the conversational context are different from the generally evaluated news domain. Hence, designing robust metrics that can capture these errors is very important. Our work presents a step in that direction by systematically benchmarking existing automatic MT metrics on machine-translated chats. Given the highly contextual nature of the chat domain, we extend and evaluate context-based reference-free and reference-based metrics.

Best Practices Given our findings, we recommend the community adopt the following practices:

1. **When references are available**: As COMET-22 achieved the best correlation across the board among all the settings, we recommend using COMET-22 as the primary evaluation metric in this scenario. We further recommend using error-predicting metrics like METRICX-23-XL or XCOMET-XL for finer-grained analysis of the severity of the errors.

2. **When references are unavailable**: In the more realistic scenario where references are unavailable, we recommend using COMET-20-QE *as is* for evaluating translations in into English and the context-aware COMET-20-QE for evaluating translations in out of English settings respectively at the turn level. For dialogue-level evaluation, SLIDE would be the most reliable among existing alternatives.

Alternatively, our proposed reference-free CONTEXT-MQM metric based on GPT-4, can also be used as it outperformed reference-based COMET-22 on the small-scale preliminary evaluation (§ 5.3). Though, in the light of recent studies showing potential biases with LLM-based evaluation – LLMs might favor their own outputs – we recommend using LLM-based evaluation only in scenarios where the evaluating LLM does not generate the translations (Panickssery et al., 2024).

Is Context Useful and Under What Conditions? Our experiments and analysis shed some light on how and when the context can be helpful. Notably, we show that context adds little information in the presence of a reference translation or when evaluating translation quality into English. However, adding context improves quality assessment across error types for reference-free evaluation in out-of-English settings, especially when it provides useful information for ambiguity resolution and is correct and complete. However, there remain several open questions and directions for future work:

1. *Improving Detection of MT errors:* As illustrated by our results, although COMET-22 achieves high correlations with human judgments, there is a drop in correlation for imperfect segments, suggesting a need for designing a metric that can do well at estimating quality for both perfect and imperfect translations.

2. *Better Reference-free Evaluation:* Our findings show that reference-free learned metrics lag behind reference-based ones in evaluating the translation quality of bilingual chats. This presents opportunities to develop effective evaluation methods lacking reference translations.

3. *Optimizing Context Utilization:* We implemented a simple approach to utilize context from both participants in both learned metrics and for LLM-based MT evaluation. However, it remains to be investigated how context can be utilized in other ways and when the metric should rely on the contextual information.

Our work, hence, opens avenues for integrating context-based signals in chat quality assessment and chat translation as well as paves the way for a more finer-grained analysis of the type, nature, and selection of contextual signals.

Acknowledgments

We thank Ben Peters, António Farinhas, and Duarte Alves for their useful and constructive comments. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*.

Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.52

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4755

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2302

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3346

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. https://doi.org/10.18653/v1/2021.naacl-main.280

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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. https://doi.org/10.18653/v1/2020.acl-main.747

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.798

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ana C. Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023a. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.100

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023b. When

does translation require context? A data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.36

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.505

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5401

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. https://doi.org/10.18653/v1/2023.wmt-1.51

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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is better than one: Improving multilingual collaboration by giving two machine translation outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 852–863, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2675133.2675197

Madalena Gonçalves, Marianna Buchicchio, Craig Stewart, Helena Moniz, and Alon Lavie. 2022. Agent and user-generated content and its impact on customer support MT. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 201–210, Ghent, Belgium. European Association for Machine Translation.

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

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *ArXiv*, abs/1704.05135.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.111

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.63

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.64

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. QUality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3037

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. 2022. Chat translation error detection for assisting cross-lingual communications. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 88–95, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463. https://doi.org/10.5565/rev/tradumatica.77

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. 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.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3336

André F. T. Martins, Joao Graca, Paulo Dimas, Helena Moniz, and Graham Neubig. 2020. Project MAIA: Multilingual AI agent assistant. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 495–496, Lisboa, Portugal. European Association for Machine Translation.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6311

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1313

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Miguel Menezes, M. Amin Farajian, Helena Moniz, and João Varelas Graça. 2023. A context-aware annotation framework for customer support live chat machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 286–297, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

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. https://doi.org/10.3115/1073083.1073135

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. https://doi.org/10.18653/v1/W15-3049

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Evaluating metrics for document-context evaluation in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 812–814, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.68

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020a. 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, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, 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. 2022b. 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.

Samantha Robertson and Mark Díaz. 2022. Understanding and being understood: User strategies for identifying and recovering from mistranslations in machine translation-mediated chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2223–2238. https://doi.org/10.1145/3531146.3534638

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. https://doi.org/10.18653/v1/2020.acl-main.704

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.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3031

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *DiscoMT@EMNLP*. https://doi.org/10.18653/v1/W17-4811

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1116

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 679–688. https://doi.org/10.1145/1518701.1518807

Jerrold H. Zar. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7. https://doi.org/10.1002/0470011815.b2a15150

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating

with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1053

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains.