Improving Context Usage for Translating Bilingual Customer Support Chat with Large Language Models

José Pombal[1,2,3*], **Sweta Agrawal**[2*], **André F.T. Martins**[1,2,3,4]
[1]Unbabel [2]Instituto de Telecomunicações
[3]Instituto Superior Técnico, Universidade de Lisboa [4]ELLIS Unit Lisbon
jose.pombal@unbabel.com,swetaagrawal20@gmail.com

Abstract

This paper describes Unbabel+IT's submission to the Chat Shared Task held at the Workshop of Machine Translation 2024. The task focuses on translating customer support chats between agents and customers communicating in different languages. We present two strategies for adapting state-of-the-art language models to better utilize contextual information when translating such conversations. Our training strategy involves finetuning the model on chat datasets with context-augmented instructions, resulting in a specialized model, TOWERCHAT. For inference, we propose a novel quality-aware decoding approach that leverages a context-aware metric, CONTEXTCOMET, to select the optimal translation from a pool of candidates. We evaluate our proposed approach on the official shared task datasets for ten language pairs, showing that our submission consistently outperforms baselines on all and competing systems on 8 out of 10 language pairs across multiple automated metrics. Remarkably, TOWERCHAT outperforms our contrastive submission based on the much larger TOWER-V2-70B model while being 10× smaller. According to human evaluation, our system outperforms all other systems and baselines across all language pairs. These results underscore the importance of context-aware training and inference in handling complex bilingual dialogues.

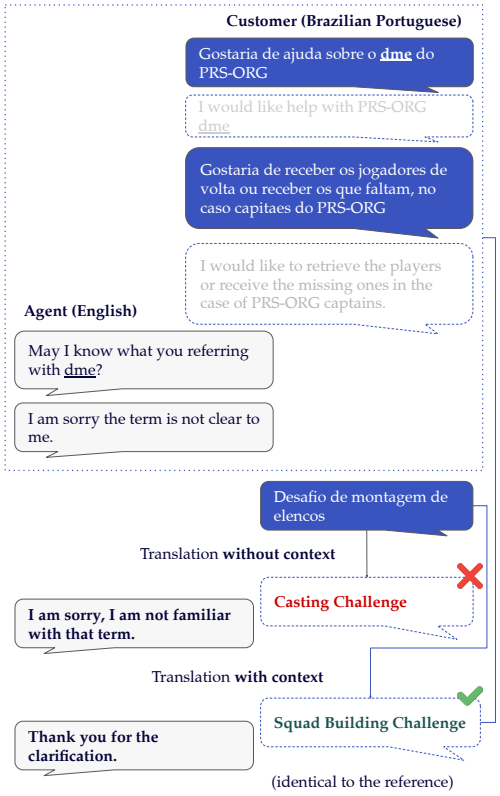

Figure 1: A WMT24 sample conversation (some turns omitted) with reference English translations. Without context, TOWERCHAT mistranslates "montagem de elencos" to "casting". With context, it correctly translates the source, understanding the customer is talking about a squad building challenge ("dme").

1 Introduction

The focus of this year's chat translation (Chat MT) shared task is the translation of conversations in customer service applications. This task differs from classical MT in that the interactions are bilingual and the texts are often more dynamic, contextualized, and informal than the structured content typically found in news or Wikipedia articles. In such scenarios, leveraging conversation context could potentially help avoid cases of lexical inconsistency and incoherence (Läubli et al., 2018; Toral et al., 2018). However, previous editions of the Chat MT shared task have shown that standard MT models are still incapable of doing so (Farajian et al., 2020; Farinha et al., 2022).

Large Language Models (LLMs) present a promising avenue to address this issue. Not only are they becoming the state-of-the-art solution for multilingual machine translation (Zhang et al., 2023; Wei et al., 2023; Alves et al., 2023; Reinauer et al., 2023; Zhu et al., 2024), but they are also known to handle context adeptly (Karpinska and

*Equal contribution.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 993–1003
November 15–16, 2024.

Iyyer, 2023; Wang et al., 2023; He et al., 2024). Despite their potential, the application of LLMs in understanding and translating bilingual chat conversations remains underexplored. We aim to bridge this gap by investigating how translation LLMs can be adapted for the Chat MT task and how they can effectively leverage conversational context to produce better translations.

Using TOWER LLM (Alves et al., 2024), a strong LLM specialized for MT and related tasks, we show that an LLM not fine-tuned for the chat domain struggles to leverage context for disambiguation, often resulting in translations that are worse than those produced without context. We thus propose two solutions to improve context usage for translation LLMs. First, we build a translation model tailored for Chat MT – TOWERCHAT – finetuned on a carefully constructed context-augmented dataset. Second, to further improve the usage of contextual information during inference, we take a novel approach of performing quality-aware decoding (Fernandes et al., 2022, QAD) with a context-aware MT evaluation metric, CONTEXTCOMET (Agrawal et al., 2024). QAD approaches select one best hypothesis from a pool of candidates using an MT metric, and have been shown to improve translation quality (Freitag et al., 2022; Fernandes et al., 2022; Farinhas et al., 2023).

This serves as our primary submission to the WMT24 Chat MT shared task, along with two contrastive ones – TOWERCHAT without QAD, and TOWER-V2-70B. The TOWER-V2-70B model is the strongest version of TOWER, which was developed for the General MT shared task.[1] The translations obtained from our approach consistently achieve the best scores across all language pairs tested, as measured by both automatic MT metrics (neural and lexical) and lexical cohesion metrics (MUDA accuracy) and human evaluation, beating strong baselines that disregard the context of conversations. Furthermore, TOWERCHAT without QAD maintains general translation capabilities and achieves better or comparable quality to TOWER-V2-70B, outlining the importance of in-domain adaptation of translation LLMs on Chat MT data.

2 Chat Translation Shared Task: Dataset and Challenges

This year's chat MT dataset includes bilingual online customer service chats between an English-speaking agent and clients who speak Portuguese, French, Italian, Dutch, or Korean. These conversations are often unplanned, informal, and nonstandard, contrasting with the well-formed text of most other translation domains. An example conversation is shown in Figure 1.

We present the general statistics from this year's shared task datasets in Table 1, including (i) the number of instances in the dataset for each language pair; (ii) the average character length of the source segments; (iii) the average number of segments in a conversation and (iv) the percentage of segments tagged with MUDA (Fernandes et al., 2023), an automatic tagger for identifying tokens belonging to certain discourse classes (lexical cohesion, verb forms, pronouns, formality) of potentially ambiguous translations. While the development and test sets exhibit a similar distribution in terms of segment length and count, they differ significantly from the training dataset. Furthermore, up to 30% en↔fr instances are tagged as requiring disambiguation according to MUDA, highlighting the complexity and the need for contextual information to generate high-quality translations.[2]

Next, we describe the process of building TOWERCHAT, which was conditioned by the aforementioned inherent complexities of Chat MT.

3 Adapting TOWER for Chat Translation

LLMs have shown the potential to use contextual information to perform many NLP tasks (Karpinska and Iyyer, 2023). In this work, we investigate whether providing contextual information can improve translation quality for bilingual chats using strong translation LLMs like TOWERINSTRUCT. Contrary to our expectations, our preliminary results indicate that incorporating context into the prompt instruction diminishes overall translation quality. We believe this is due to TOWERINSTRUCT's training data lacking chat-specific MT examples, which results in the model's unfamiliarity with the context format and the inability to adequately use context (Section 5). To mitigate this and improve the usage of contextual information, we propose two strategies – one for training and one for inference.

[1] Private model, but since we developed it, we have access.

[2] Note that MUDA only tags formality for Korean and does not detect instances of semantic ambiguity. The dataset likely features many more complex context phenomena.

Language Pair	# Instances			Avg. Source Length			Avg. # Segments per Conversation			% MuDA tagged	
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Dev	Test
en↔de	17805	2569	2041	47.40	52.26	53.09	36.12	31.33	30.46	15.65	15.78
en↔fr	15027	3007	2091	41.84	54.90	56.23	56.92	33.41	32.17	29.43	29.65
en↔pt-br	15092	2550	2040	42.72	46.46	46.49	34.69	26.56	27.95	13.02	12.99
en↔ko	16122	1935	1982	39.86	47.67	46.90	38.11	50.92	47.19	0.41	0.50
en↔nl	15463	2549	2015	45.40	52.31	54.31	25.99	35.40	34.74	22.01	23.13

Table 1: Statistics for each language pair and split of the data for the WMT24 Chat MT shared task.

Context: {context}
Translate the following **{source_lang}** source text to **{target_lang}**, given the context.
{source_lang}: **{source_seg}**
{target_lang}: {target_seg}

Figure 2: Instructions with context for Chat MT. Parts in purple are only included when a context is available.

3.1 Finetuning on Context-augmented Chats

For a conversation C of length L with segments $\{(x_t, y_t, c_t)\}_{i=1}^{L}$, where x_t is a text generated by the agent or the customer, y_t is its reference translation in the target language, and c_t is the preceding bilingual context, we train the model to minimize the cross-entropy loss using the input prompt shown in Figure 2:

$$\mathcal{L} = -\log P(y_t|x_t, c_t). \quad (1)$$

The context c_t includes all previous turns of the conversation, capturing important discourse-level information such as pronoun references, formality, and other pragmatic elements that influence the translation. For the first turn, no context is available, so the prompt reduces to the standard format used for zero-shot MT, as described in Alves et al. (2024). We train TOWERCHAT by finetuning TOWERBASE 7B on the concatenation of TOWERBLOCKS and the entire training dataset of the shared task, using context-aware prompts. This endows the model with the capacity to better understand and leverage conversational context, enabling it to generate high-quality translations.

3.2 QAD with Context-aware Metrics

Decoding strategies informed by translation quality metrics such as Minimum Bayes Risk Decoding (MBR) and Tuned Reranking (TRR) have been shown to consistently improve output quality over greedy decoding (Fernandes et al., 2022; Freitag et al., 2022; Nowakowski et al., 2022; Farinhas et al., 2023). In QAD, the goal is to find a translation among a set of candidates that maximizes the expected utility function, often measured using an MT metric like reference-based COMET. Recently, Agrawal et al. (2024) showed that context-aware MT metrics correlate better with human judgments compared to their non-contextual counterparts, especially when evaluating out-of-English chat translations. The context-aware versions of COMET (Vernikos et al., 2022; Agrawal et al., 2024) compute quality scores for a source-reference-hypothesis tuple, $(x, y, \hat{y})$, using the representations extracted from a context-augmented input, $([c; x], [c; y], [c; \hat{y}])$.

As such, we use CONTEXTCOMET for MBR decoding in our submission. For a given source text x, the previous bilingual context, c, and a set of candidate translations sampled from the model $\mathcal{Y}$, the utility of each candidate $\hat{y} \in \mathcal{Y}$, is given by

$$u = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \text{CONTEXTCOMET}([c; x], [c; y], [c; \hat{y}]). \quad (2)$$

The best translation is selected using:

$$y_{\text{mbr}} := \arg\max_{\hat{y} \in \mathcal{Y}} [u(\hat{y}, \mathcal{Y})]. \quad (3)$$

This enables the model to select a translation amongst alternative hypotheses, potentially leading to more accurate and contextually appropriate outputs. QAD with TOWERCHAT serves as our primary submission to the Chat Shared Task.

4 Experimental Configurations

Baselines. We report the shared task's official baseline: NLLB-3.3B with beam search decoding (beam width: 4). Additionally, we report greedy decoding results with TOWERINSTRUCT-7B, and TOWER-V2-70B, the strongest TOWER model. The former serves as a direct baseline for our method, while the latter is a state-of-the-art baseline for MT.

TOWERCHAT. We report greedy and QAD results with the TOWERCHAT-7B model. For

Model	EN-XX				XX-EN			
	chrF↑	COMET↑	METRICX↓	CONTEXT-QE↑	chrF↑	COMET↑	METRICX↓	CONTEXT-QE↑
Baselines								
NLLB	59.78 9	88.61 8	1.04 6	4.95 6	70.76 9	88.16 7	0.74 5	5.06 6
TOWERINSTRUCT 7B (0-shot)								
w/o context	64.95 8	91.69 6	0.38 3	16.29 4	76.04 6	92.17 5	0.56 4	15.73 4
w/ context	63.39 9	91.09 7	0.49 5	14.53 5	74.32 8	91.36 6	0.60 4	14.67 4
TOWERINSTRUCT 7B (5-shot)								
w/o context	65.20 8	91.75 6	0.39 3	16.62 3	75.84 7	92.22 5	0.54 4	15.97 3
w/ context	63.62 9	91.03 7	0.50 5	15.02 4	73.52 9	91.64 6	0.59 4	14.67 4
TOWER-V2 70B (5-shot)								
w/o context	68.26 5	92.68 4	0.30 2	18.24 2	77.17 4	92.69 3	0.47 2	17.71 1
w/ context	68.26 6	92.50 4	0.30 2	17.53 2	76.03 6	92.37 4	0.46 2	17.28 2
TOWERCHAT								
w/o context	71.68 5	93.01 4	0.32 3	16.77 3	77.97 3	92.72 4	0.51 3	16.40 3
w/ context	75.93 3	93.63 3	0.32 3	16.61 3	78.87 2	93.01 3	0.47 2	16.15 3
+ QAD (COMET)	76.36 2	**94.18** 1	**0.25** 2	**18.78** 1	**78.92** 2	**93.39** 1	**0.44** 1	18.18 1
+ QAD (CONTEXTCOMET)	**76.56** 1	94.05 2	0.26 2	18.68 1	**78.92** 2	93.24 2	**0.44** 1	**18.24** 1

Table 2: Main Results on Official Test Set: QAD with TOWERCHAT outperforms all baselines across the board. Models with statistically significant performance improvements are grouped in quality clusters

QAD, we perform MBR with COMET or CONTEXTCOMET on 100 candidates obtained via epsilon sampling with $\epsilon = 0.02$ (Hewitt et al., 2022).

Instruction settings. To assess whether systems can properly leverage conversational context, we prompt the LLM-based MT with two instruction formats (see Figure 2): 1) **w/o context**, where the model is prompted without any conversational context (without the purple highlighted text). 2) **w/ context**, where the entire previous bilingual conversation is provided as the context in the prompt.[3]

Evaluation. We report the final results on the shared task's test set on all ten language pairs. As exemplified in Figure 1, ambiguous contextual phenomena often arise in Chat MT that require nuanced evaluation. As such, we leverage three types of assessments: 1) automatic metrics for measuring overall translation quality – two neural and one lexical – COMET-22 (Rei et al., 2022), CHRF (Popović, 2015) and METRICX-XL (Juraska et al., 2023); 2) a reference-free neural metric that uses context for quality assessment, CONTEXT-QE (Agrawal et al., 2024); 3) F1-score on MUDA tags for measuring whether models correctly resolve lexical ambiguities (Fernandes et al., 2023). Considering METRICX, CHRF, and MUDA is crucial in our case, as COMET may favor the QAD strategies we use.

On Tables 2 and 3 we report performance clusters based on statistically significant performance gaps at a 95% confidence threshold.[4] We create per-language groups for systems with similar performance, following Freitag et al. (2023), and obtain system-level rankings using a normalized Borda count (Colombo et al., 2022), which is defined as an average of the obtained clusters. Note that a first cluster will not exist if no model significantly outperforms all others on a majority of languages.

5 Main Results

Table 2 presents the average results for EN→XX and XX→EN translation directions. TOWERCHAT with QAD outperforms all baselines across all settings on automatic metrics and human evaluation.

TOWERCHAT leverages context more adeptly than TOWERINSTRUCT. Our primary goal in this task was to create a model that can effectively leverage context to generate high-quality translations with LLMs. As shown in Table 2, TOWERCHAT consistently outperforms TOWERINSTRUCT across all settings, language pairs and evaluation metrics. Furthermore, TOWERCHAT shows an average improvement of 4 CHRF points for en-xx when using context (*w/ context*), compared to a context-agnostic prompt (*w/o context*).[5] This trend also holds when evaluating translation quality using the primary metric, COMET, for 8 out of 10 language

[3]Note that **{target_seg}** is unavailable during inference and the model is asked to perform prompt completion.

[4]For segment-level metrics, such as COMET, we perform significance testing at the segment level. For CHRF, we substitute segment-level scores with corpus-level scores calculated over 100 random samples, each with a size equal to 50% of the total number of segments.

[5]The improvement is statistically significant with a 92.1% accuracy (Kocmi et al., 2024).

Model	EN-XX					XX-EN				
	DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines										
NLLB	90.56 7	91.06 6	86.33 9	87.26 9	87.86 8	89.03 6	89.18 6	86.1 8	88.05 9	88.45 8
TOWERINSTRUCT 7B (0-shot)										
w/o context	91.71 5	91.89 5	91.9 7	91.64 5	91.3 7	92.08 4	92.78 2	90.43 7	93.13 6	92.45 5
w/ context	91.48 6	91.08 6	90.79 8	91.13 7	91.0 7	91.33 5	91.89 5	90.63 6	91.88 7	91.08 7
TOWERINSTRUCT 7B (5-shot)										
w/o context	91.75 5	91.75 5	92.32 6	91.41 6	91.55 6	92.06 4	92.28 4	90.63 6	93.55 5	92.6 5
w/ context	91.41 6	90.88 6	90.85 8	90.45 8	91.58 6	92.06 4	92.14 5	90.82 5	90.89 8	92.29 6
TOWER-V2 70B (5-shot)										
w/o context	92.81 2	92.21 4	93.06 5	92.55 4	92.76 5	**92.68** 1	**93.23** 1	91.46 4	93.08 6	92.98 3
w/ context	92.61 3	92.08 4	93.03 5	91.76 5	93.02 4	92.07 4	92.44 4	91.42 4	93.05 6	92.89 3
TOWERCHAT										
w/o context	92.36 4	92.26 4	93.89 4	93.73 3	92.81 4	92.28 3	92.79 2	91.06 5	94.69 4	92.78 4
w/ context	92.74 2	92.64 3	94.53 3	94.16 2	94.09 3	92.24 3	92.67 3	92.09 3	94.98 3	93.06 3
said + QAD (COMET)	**93.28** 1	**93.13** 1	**94.91** 1	**95.01** 1	**94.54** 1	92.58 1	92.95 2	**92.63** 1	**95.32** 1	**93.49** 1
+ QAD (CONTEXTCOMET)	93.22 1	92.96 2	94.76 2	94.96 1	94.36 2	92.48 2	92.71 3	92.46 1	95.16 2	93.38 2
Official Rank (COMET)	2nd	1st	1st	1st	1st	2nd	1st	1st	1st	1st
Official Rank (Human)	1st	1st	1st	1st	1st	1st	1st	1st	1st	1st

Table 3: Main Results by COMET on Official Test Set by Language Pair. Models with statistically significant performance improvements are grouped in quality clusters.

Model	Lexical Cohesion	Verb Form	Pronouns	Formality
NLLB	72.43	52.97	72.31	56.44
TOWERCHAT	85.13	47.80	79.71	81.93
QAD (COMET)	85.94	61.22	**80.56**	82.46
QAD (CONTEXTCOMET)	**86.21**	**64.38**	79.28	**83.16**

Table 4: MuDA F1 results. On average, QAD with CONTEXTCOMET has the best F1 score.

Models	EN→XX	XX→en
TOWERINSTRUCT-7B	84.28	82.77
TOWERCHAT-7B	83.95	82.54

Table 5: COMET scores for TOWERINSTRUCT and TOWERCHAT on the WMT23 test set.

pairs as shown in Table 3. We attribute this to the inclusion of context-augmented Chat MT instruction dataset in TOWERCHAT's training, highlighting the effectiveness of in-domain fine-tuning.

QAD results in consistent gains over greedy decoding, surpassing 70B models. The highest-quality translations according to all metrics considered are obtained after performing QAD with COMET or CONTEXTCOMET on top of TOWERCHAT-7B, even outperforming the much larger TOWER-V2-70B, which uses few-shot examples. Moreover, QAD closes the gap in quality as measured by METRICX and CONTEXT-QE between TOWERCHAT-7B (greedy) and TOWER-V2-70B models, demonstrating that advanced inference techniques can effectively make smaller models competitive against much larger ones.

Context-aware QAD improves MUDA F1 over Context-agnostic QAD. While all neural and lexical metrics indicate that QAD with CONTEXTCOMET and COMET perform comparably, these metrics may not fully capture nuanced differences in translation quality. To address this, we evaluate MUDA F1 accuracy scores for a subset of models in Table 4. The results show that QAD with CONTEXTCOMET consistently outperforms QAD with COMET across all dimensions, except pronouns. Our qualitative analysis suggests that the pronoun accuracy might have been lower due to potential paraphrasing. Coupled with the previous results, these findings strongly motivate further exploration of QAD with context-aware metrics.

Finetuning on Chat data does not degrade general translation capabilities. To ensure that adding chat MT dataset in the mix does not impact the generic translation capabilities of LLMs, we report COMET on the standard WMT23 benchmark (Kocmi et al., 2023) averaged across EN→XX and XX→EN directions for TOWERINSTRUCT and TOWERCHAT in Table 5. TOWERCHAT suffers only minor degradation (−0.3) relative to TOWERINSTRUCT, validating the viability and effectiveness of our finetuning approach.

System	EN-DE			EN-FR			EN-NL			EN-PT			EN-KO		
	T (XX)	T (EN)	C	T (XX)	T (EN)	C	T (XX)	T (EN)	C	T (XX)	T (EN)	C	T (XX)	T (EN)	C
Baseline	78.05	87.57	74.50	80.59	77.82	67.81	82.66	90.98	53.07	61.27	73.98	56.37	79.13	90.47	85.63
Unbabel-IT	89.42	92.74	84.22	90.24	90.00	79.62	98.16	97.40	92.22	82.04	82.37	78.00	93.39	96.31	93.21

Table 6: Human Evaluation results on the official test set. T and C represent aggregated turn-level and conversation-level direct-assessment scores respectively.

6 Human Evaluation

TOWERCHAT is the winner of the WMT24 Chat MT Shared Task across all language pairs according to human evaluation. Table 6 shows that our model significantly surpasses the baseline on both turn-level (T) and conversation-level (C) evaluations in all language directions. Notably, it reaches an average direct assessment score of > 90 at both turn-level and conversation-level for EN-FR, EN-NL, and EN-KO translation pairs. The victory on conversation-level evaluation outlines the superior capacity of TOWERCHAT to incorporate bilingual conversational context when translating.

That said, there is a visible drop between turn-level and conversation-level scores, leaving room for improvement on how well TOWERCHAT leverages context for translation. In future work, we wish to explore thoroughly under what circumstances context is useful to produce a better translation, and to what extent TOWERCHAT can leverage it appropriately.

7 Conclusion

In this work, we present two strategies for improving context usage for bilingual chat translation using LLMs. Our training strategy involves finetuning LLMs on context-augmented instructions resulting in higher-quality translations during inference when using bilingual context. Second, we propose a novel quality-aware decoding strategy with a context-aware metric (CONTEXTCOMET) that significantly improves translation quality across the board, surpassing a state-of-the-art 70B translation model and all other baselines. Our findings show successful usage of contextual information as measured by MUDA in resolving ambiguities for the highly contextual domain of chat translation. Crucially, our system finished first in human evaluation across all the shared task's language pairs.

Acknowledgments

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

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André FT Martins. 2024. Is context helpful for chat translation evaluation? *arXiv preprint arXiv:2403.08314*.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? new perspectives on nlp benchmarking. In *Advances in Neural Information Processing Systems*.

M Amin Farajian, António V Lopes, André FT Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the wmt 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75.

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.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

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*, Singapore. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. *arXiv preprint arXiv:2401.06760*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334, Abu Dhabi, United Arab Emirates (Hybrid). 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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. 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.

Raphael Reinauer, Patrick Simianer, Kaden Uhlig, Johannes E. M. Mosig, and Joern Wuebker. 2023. Neural machine translation models can learn to be few-shot learners. *Preprint*, arXiv:2309.08590.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

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.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *Preprint*, arXiv:2306.10968.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Validation Results

MODEL	EN-XX		XX-EN	
	CHRF	COMET	CHRF	COMET
Baselines				
NLLB 3.3B	58.41	86.97	65.39	85.51
TOWERINSTRUCT 7B (0-shot)				
w/o context	63.69	90.69	71.57	90.62
w/ context	63.51	90.53	70.16	89.84
TOWER-V2 70B (5-shot)				
w/o context	67.08	91.95	73.41	91.41
w/ context	66.85	91.69	71.87	90.94
TOWERCHAT				
w/o context	70.63	92.21	73.42	91.13
w/ context	74.17	92.76	73.81	91.35
+ QAD (COMET)	74.49	**93.49**	73.93	**91.85**
+ QAD (CONTEXTCOMET)	**74.54**	93.31	**74.15**	91.70

Table 7: Results on the Validation Set: TOWERCHAT with QAD outperforms all baselines.

B Test Results by Language Pair

MODEL	EN-XX					XX-EN				
	DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines										
NLLB	70.22	76.03	58.60	34.50	59.55	71.79	76.37	67.13	69.87	68.62
TOWERINSTRUCT 7B (0-shot)										
w/o context	71.81	74.59	72.26	43.18	62.90	77.57	79.02	72.06	75.73	75.80
w/ context	71.16	74.38	68.50	41.70	61.23	75.68	78.31	71.83	72.63	73.15
TOWERINSTRUCT 7B (5-shot)										
w/o context	71.38	74.72	72.59	42.76	64.55	76.64	78.67	71.68	76.23	75.96
w/ context	71.48	73.66	66.15	40.94	65.86	75.05	77.56	70.39	70.87	73.74
TOWER-V2 70B (5-shot)										
w/o context	75.58	75.53	75.02	47.16	68.00	78.07	80.49	73.58	76.57	77.12
w/ context	74.60	75.28	74.05	46.99	70.38	77.54	77.63	73.16	75.69	76.10
TOWERCHAT										
w/o context	74.04	77.12	79.71	57.63	69.91	79.31	79.36	74.00	80.17	77.01
w/ context	76.41	79.97	82.24	61.27	79.78	79.91	79.26	75.72	81.30	78.15
+ QAD (COMET)	77.09	80.34	82.25	61.79	80.33	79.70	78.78	75.88	81.56	78.67
+ QAD (CONTEXTCOMET)	77.23	80.51	82.55	62.29	80.25	79.87	78.57	76.01	81.57	78.60

Table 8: Results by CHRF (higher is better) on Official Test Set by Language Pair.

MODEL	EN-XX					XX-EN				
	DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines										
NLLB	0.62	0.38	1.57	1.51	1.13	0.65	0.70	1.07	0.59	0.68
TOWERINSTRUCT 7B (0-shot)										
w/o context	0.28	0.23	0.43	0.57	0.37	0.50	0.53	0.86	0.37	0.53
w/ context	0.38	0.29	0.69	0.60	0.49	0.56	0.55	0.74	0.46	0.69
TOWERINSTRUCT 7B (5-shot)										
w/o context	0.29	0.25	0.39	0.62	0.39	0.50	0.55	0.79	0.37	0.52
w/ context	0.35	0.32	0.69	0.75	0.39	0.54	0.59	0.72	0.60	0.51
TOWER-V2 70B (5-shot)										
w/o context	0.24	0.22	0.27	0.45	0.30	0.48	0.46	0.63	0.33	0.45
w/ context	0.25	0.21	0.28	0.45	0.29	0.50	0.48	0.58	0.31	0.42
TOWERCHAT										
w/o context	0.27	0.24	0.29	0.42	0.37	0.50	0.51	0.71	0.33	0.52
w/ context	0.34	0.26	0.27	0.45	0.27	0.47	0.48	0.60	0.30	0.48
+ QAD (COMET)	0.30	0.22	0.24	0.31	0.21	0.46	0.46	0.55	0.27	0.45
+ QAD (CONTEXTCOMET)	0.31	0.22	0.24	0.29	0.23	0.47	0.47	0.56	0.27	0.45

Table 9: Results by METRICX (lower is better) on Official Test Set by Language Pair.

MODEL	EN-XX					XX-EN				
	DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines										
NLLB	15.56	1.24	-5.51	4.11	9.35	19.09	0.77	-6.75	4.13	8.04
TOWERINSTRUCT 7B (0-shot)										
w/o context	21.84	8.96	9.11	19.73	21.83	23.41	7.46	7.49	18.66	21.64
w/ context	21.26	7.22	6.89	17.50	19.79	22.52	7.45	7.12	17.72	18.53
TOWERINSTRUCT 7B (5-shot)										
w/o context	21.75	9.41	10.47	19.50	21.95	23.37	8.29	8.51	18.35	21.33
w/ context	21.83	8.20	8.17	15.22	21.68	22.93	6.75	7.16	15.49	21.02
TOWER-V2 70B (5-shot)										
w/o context	23.42	10.47	12.38	20.84	24.07	25.21	9.77	10.26	20.08	23.21
w/ context	23.13	9.74	12.30	18.91	23.56	25.11	9.91	9.72	18.88	22.80
TOWERCHAT										
w/o context	22.31	9.15	10.55	20.08	21.75	24.12	7.72	8.81	19.48	21.85
w/ context	22.39	8.69	11.36	18.58	22.05	24.28	7.45	9.06	17.96	21.97
+ QAD (COMET)	24.27	10.92	13.01	21.65	24.04	26.12	9.67	10.77	21.02	23.31
+ QAD (CONTEXTCOMET)	24.41	10.67	12.74	21.64	23.93	26.15	10.00	10.59	21.08	23.39

Table 10: Results by CONTEXT-QE (higher is better) on Official Test Set by Language Pair.

C MUDA F1 Scores by Language Pair

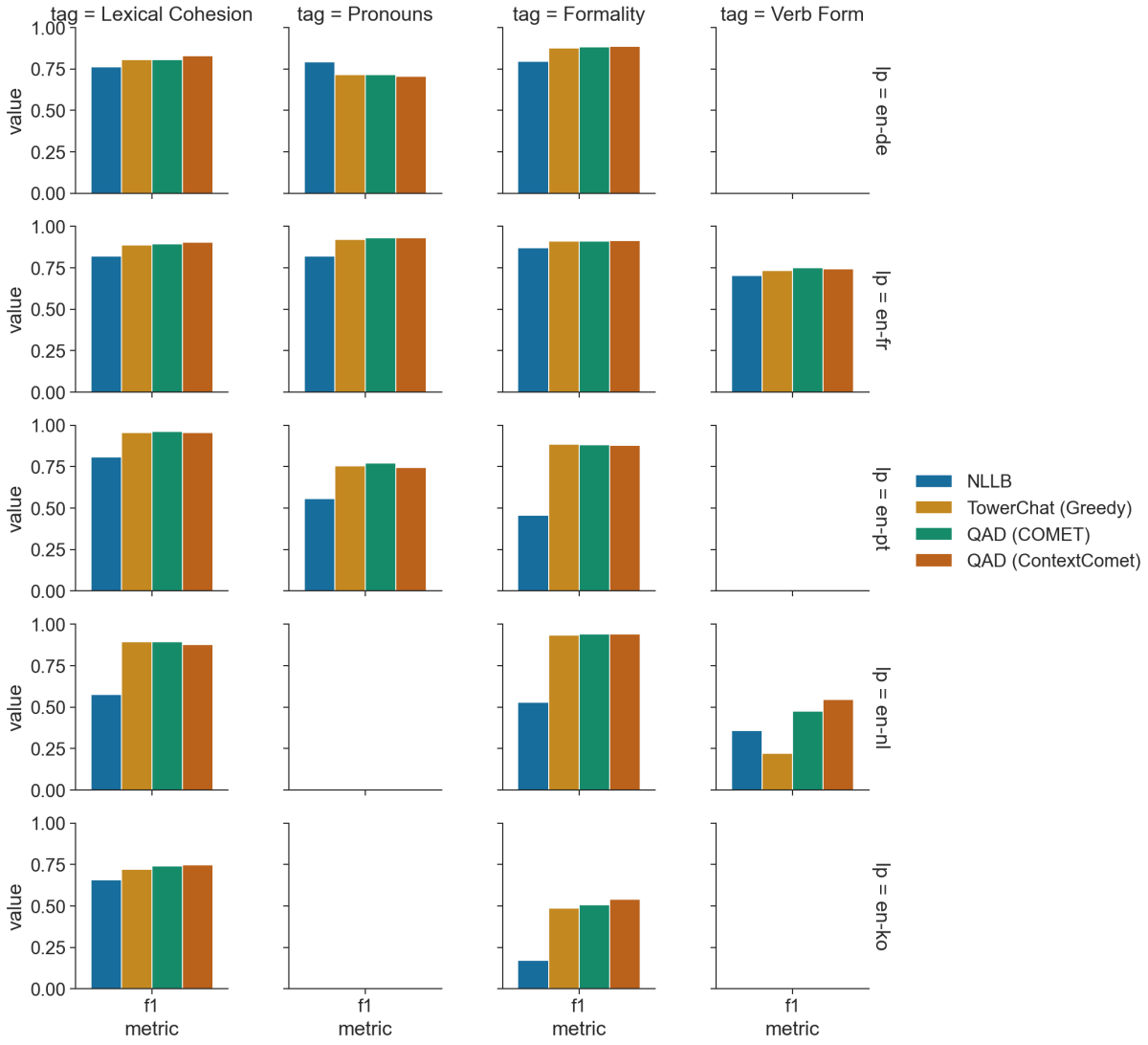

Figure 3: MuDA accuracy scores by LPs. Plots are left empty for the cases MuDA does not return tags (e.g., verb form for Korean).