A Context-aware Framework for Translation-mediated Conversations

José Pombal[1,2,3*], **Sweta Agrawal**[2*]
Emmanouil Zaranis[2,3], **Patrick Fernandes**[2,3,4], **André F. T. Martins**[1,2,3,5]

[1]Unbabel, Portugal [2]Instituto de Telecomunicações, Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[4]Carnegie Mellon University, USA [5]ELLIS Unit Lisbon, Portugal
jose.pombal@unbabel.com, swetaagrawal20@gmail.com

Abstract

Automatic translation systems offer a powerful solution to bridge language barriers in scenarios where participants do not share a common language. However, these systems can introduce errors leading to misunderstandings and conversation breakdown. A key issue is that current systems fail to incorporate the rich contextual information necessary to resolve ambiguities and omitted details, resulting in literal, inappropriate, or misaligned translations. In this work, we present a framework to improve large language model-based translation systems by incorporating contextual information in bilingual conversational settings during training and inference. We validate our proposed framework on two task-oriented domains: customer chat and user-assistant interaction. Across both settings, the system produced by our framework—TowerChat—consistently results in better translations than state-of-the-art systems like GPT-4o and TowerInstruct, as measured by multiple automatic translation quality metrics on several language pairs. We also show that the resulting model leverages context in an intended and interpretable way, improving consistency between the conveyed message and the generated translations.[1]

1 Introduction

In today's globalized world, the demand for efficient cross-lingual communication has surged across diverse domains, whether it be for providing global customer support (DePalma et al., 2006; Zhang and Misra, 2022), for enabling real-time multilingual collaboration in meetings (Zhang et al., 2021, 2022), or for facilitating effective patient-doctor interactions (Mehandru et al., 2023). This need extends beyond human-to-human communication to human-machine interactions, where LLMs have emerged as powerful tools in English but with lackluster performance in other languages (Hu et al., 2023a; Jin et al., 2024; Etxaniz et al., 2023, 2024; Liu et al., 2024; Dey et al., 2024).

One potential solution to bridge this language gap is through translation-mediated conversations. In such cases, translation serves as a middle layer between two interacting parties, be it between humans or humans and machines. In the latter case, for example, instead of relying on the model's capabilities for addressing user queries in multiple languages (i.e., direct inference), language translation and the downstream task are treated as separate problems (i.e., pretranslation) (Etxaniz et al., 2023). However, the back-and-forth nature of conversations introduces its own set of challenges, particularly in complex, multi-turn dialogues. Context can be lost, cultural nuances overlooked, and translation errors may accumulate over the conversation, leading to misunderstandings or inappropriate responses (Tsujii and Nagao, 1988; Robertson and Díaz, 2022; Mendonca et al., 2023).

Large language model (LLM)-based translation systems, however, present a promising avenue to address this issue. Not only are they becoming the state-of-the-art solution for multilingual machine translation (MT) (Zhang et al., 2023b; Wei et al., 2023; Alves et al., 2023; Reinauer et al., 2023; Zhu et al., 2024; Kocmi et al., 2023, 2024), but they are also known to handle context adeptly (Karpinska and Iyyer, 2023; Wang et al., 2023; He et al., 2024). Despite their potential, using LLMs to facilitate real-time translation-mediated conversations remains underexplored.

To tackle this problem, we propose a context-aware framework designed to enhance the

[*] Equal contribution.

[1]The datasets, outputs and the code to reproduce the findings can be found here: https://github.com/zeppombal/context-aware-mt.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 562–587, 2026. https://doi.org/10.1162/TACL.a.639
Action Editor: Masaaki Nagata. Submission batch: 11/2024; Revision batch: 2/2025; Published 4/2026.

translation capabilities of LLMs in conversation settings. During training, we use carefully constructed context-augmented examples, incorporating the original bilingual messages exchanged between the two interacting parties as context. This allows the model to attend to both contextual cues and language-specific discourse elements such as pronoun references, formality, and continuity. Additionally, we introduce quality-aware decoding (Fernandes et al., 2022, QAD) with context-aware metrics (Vernikos et al., 2022; Agrawal et al., 2024) to further help the system prioritize translations that best fit the preceding conversation.

We apply our framework in two bilingual case studies: (1) human-human conversations covering five language pairs: English $\leftrightarrow$ {German, French, Portuguese, Korean, and Dutch} and (2) human-assistant interactions in English $\leftrightarrow$ German. The assistant functions in its most proficient language, English, while users are supported in their native language via a translation layer. Our findings (§ 5) reveal that the resulting system, TowerChat, trained using context-augmented instruction training significantly improves translation quality (measured by several automatic metrics), surpassing strong baseline models (GPT-4o, TowerInstruct) in multiple language pairs. The gains extend beyond those obtained with in-domain data with standard instructions or with monolingual (English only) context. Additionally, when context is incorporated during inference with quality-aware decoding, our method reduces context-based errors in 9 out of 12 evaluated settings.

Using existing interpretability tools, we show that our model effectively uses salient parts of the context, particularly in ambiguous sentences, to generate more accurate translations. Finally, our results indicate that context is particularly beneficial when references align well with the surrounding discourse, *i.e.*, they are less surprising given the context. We also observe that the optimal context window varies across language pairs, though incorporating additional turns as context does not degrade performance. Our contributions are summarized below:

- We present a framework that integrates bilingual contextual information in LLM-based translation systems' training and inference stages when translating conversations.
- We show the efficacy of our approach in improving the accuracy of translations in multi-turn dialogues, covering human-human and human-assistant interactions.
- Our resulting system takes advantage of salient parts of the context to resolve ambiguity, improving the coherence, resolution of pronouns, and contextual accuracy of resulting translations.

2 Background

Translation often serves as a vital medium for communication when participants either do not share a common language or opt not to use it. In these situations, context plays a crucial role, directly affecting the quality and appropriateness of translations. Pronoun ambiguity, implicit references, and variations in formality present significant challenges, making accurate translation difficult without contextual cues. Context-aware MT seeks to improve translation quality by considering not just the text itself but the surrounding or broader context, which can involve linguistic, cultural, situational, or even domain-specific information. In MT research, *context* has been interpreted in various ways: the broader document or neighbouring sentences from which a source to be translated is drawn, the real-world translation setting including the intended audience, the required level of formality, or specialized terminology, among others (Castilho and Knowles, 2024).

In this work, we define *context* as **information extending beyond the current turn in bilingual, multi-turn interactions**, crucial for reducing ambiguity and maintaining coherence across turns. Unlike document-level MT, where models can process a complete document at once, translating conversations or dialogues requires turn-by-turn continuity, presenting unique challenges in maintaining consistency across exchanges.[2] We present a review of existing approaches for translating dialogues and how context has been used thus far to improve translation quality in LLMs.

[2]While document-level MT can also be framed as a form of multi-turn translation, with sentences or blocks of text functioning as individual turns, it fundamentally differs from our setting because it is neither real-time nor bilingual.

Approaches for Dialogue Translation. Recognizing the importance of context in dialogues, much prior research has focused on integrating contextual elements or additional meta-information to improve output quality. For example, Wang et al. (2016) use speaker tags for modeling the grammatical gender of the participants, while Maruf et al. (2018) incorporate conversation histories into a sentence-based attention model, improving pronoun usage and discourse coherence. Liang et al. (2021) design latent variational modules for learning the distributions of bilingual conversational characteristics (role preference, dialogue coherence, and translation consistency). Vincent et al. (2022) use extra-textual information (the speaker's gender and number of interlocutors) to improve grammatical agreement in dialogue translation. While these approaches advance dialogue translation, they rely on task-specific architectural modifications to encoder-decoder models and incorporate only limited aspects of dialogue history. In contrast, our proposed framework (§ 3) integrates the full *bilingual* contextual history seamlessly with LLMs to improve translation accuracy in dialogues.

Context-aware MT. Attempts to include extra-sentential context into the translation process date back to the pre-neural machine translation era (Webber et al., 2013) and have become increasingly common since (Maruf et al., 2021; Castilho and Knowles, 2024). However, these efforts have largely focused on cross-sentence context for document-level MT (Wang et al., 2017), where the source text is well-structured and monolingual. Moreover, most models, trained specifically for translation, have shown only marginal improvements over context-agnostic baselines (Lopes et al., 2020) and often fail to fully utilize available context (Fernandes et al., 2023).

Recently, LLMs have shown the potential to effectively use contextual information to perform many NLP tasks, including sentence and document-level translation (Karpinska and Iyyer, 2023; Wang et al., 2023). For instance, Agrawal et al. (2023), Zhang et al. (2023a), and Mu et al. (2023) retrieve relevant examples during inference and supply them as context for the current source sentence. Other approaches integrate bilingual dictionaries or domain-specific terminologies (Ghazvininejad et al., 2023; Moslem et al., 2023) or use prompts to guide LLMs in resolving ambiguity either from the given context (Pilault et al., 2023) or based on pre-existing knowledge (He et al., 2024). Additionally, Treviso et al. (2024) propose improving output quality through post-editing of initial drafts with error explanations. Wang et al. (2023) use context-aware prompts to model document-level translations during inference, whereas Wu et al. (2024) propose training LLMs with document-level context.

Despite these advances, LLMs' potential to fully exploit bilingual multi-turn contexts remains largely unexplored. Their ability to retain and utilize multilingual and cross-lingual cross-sentence context over extended text makes them well-suited for handling bilingual exchanges. To address this gap, we propose a framework that integrates bilingual context into LLMs during training and inference in conversational settings.

3 A Context-aware Framework

In this section, we outline a framework for effectively leveraging contextual information to improve translation quality in translation-mediated conversations. Our framework addresses both training and inference, showing how context can be systematically integrated to produce translations that align closely with conversational flow.

3.1 Context-augmented Instruction Fine-tuning

Translation of conversations requires understanding both the current utterance and its preceding context. Each instance may be preceded by a series of turns, which, in our case, are bilingual, as shown in Figure 2. To enable the model to effectively leverage context, we enrich the training dataset with context-augmented instructions. Specifically, for a conversation C of length L with segments $\{(x_t, y_t, c_t)\}_{t=1}^{L}$, where x_t is a text generated by a participant at turn t, y_t is its reference translation in the target language, and c_t is the relevant context, we draft a context-augmented instruction as shown in Figure 1. A training instance is shown in Figure 8.

Choice of Context. The context, c_t can be sourced from previous conversational turns, external knowledge bases, or situational factors, encapsulating crucial discourse-level information

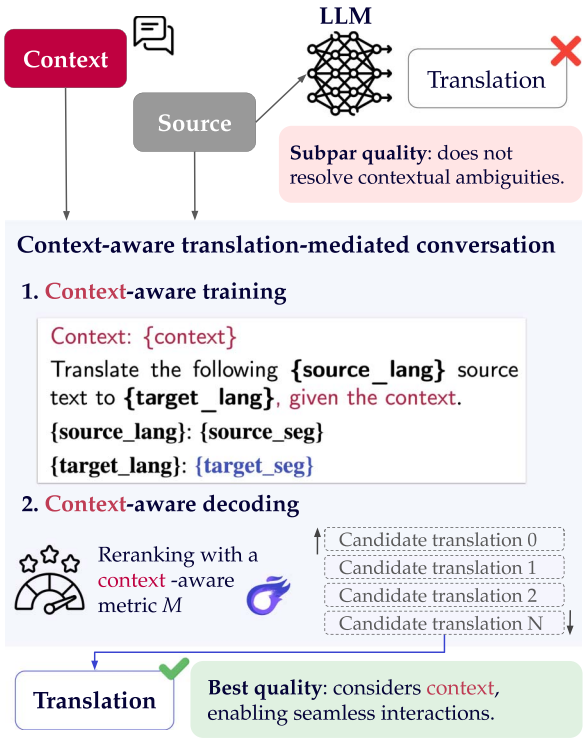


Figure 1: Our framework for optimizing LLMs for mediating conversations with translation. First, we train the LLM on a conversation translation dataset formatted with context-aware prompts. At inference time, we use a context-aware metric to select the best translation from a pool of candidates.

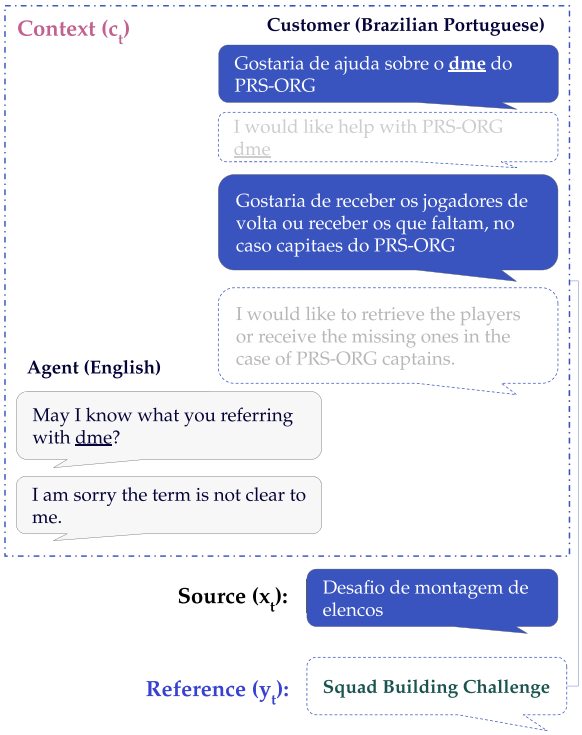


Figure 2: A sample conversation illustrating helpfulness of context in resolving ambiguity: Correctly translating the source requires inferring that the customer is referring to players (''jogadores'') in previous turns from the ''squad building challenge''.

such as pronoun references and formality. For simplicity, we include only the original *bilingual* texts from the previous turns of the participants, $x_{<t}$. This ensures the model retains a holistic view of the conversation, capturing nuances crucial for accurate translation. However, as LLMs are shown to be more robust when instructed in English (Zhao et al., 2024) and might struggle with understanding the cross-lingual context in low-resource languages, we compare our approach with one where the preceding context is provided solely in English. This requires using generated translations for non-English source texts as context during inference. We hypothesize that while cross-lingual context adds complexity, it better preserves language-specific nuances that might be lost when the context is limited to one language.

Alternative choices can involve extracting or summarizing only the relevant parts of the conversation (Krause et al., 2024; Sung et al., 2024) or incorporating generated translations along with the source texts as a part of the context (Wu et al., 2024). The former adds overhead to the pipeline and risks introducing errors or inconsistencies in the context, potentially losing critical information. Conversely, including system-generated translations during inference creates dependencies on prior outputs, which may lead to errors from inaccuracies that propagate across translations. While our framework provides flexibility in the choice of context, we leave a detailed investigation of alternatives to future work.

Training. We train the model to minimize the cross-entropy loss using a context-aware prompt:

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

This endows the model with the capacity to leverage conversational context when translating.

3.2 Quality-aware Decoding 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 improve output quality over greedy decoding (Stahlberg et al., 2017; Fernandes et al., 2022; Freitag et al., 2022; Nowakowski et al., 2022; Farinhas et al., 2023). In quality-aware

decoding, the primary goal is to find a translation among a set of candidates that maximizes an expected utility function, often measured with an automatic MT metric. While previous work has used MBR to improve the quality of individual out-of-context sentences, we extend this by incorporating context into the decoding process. To improve both output quality and contextual accuracy, we apply QAD with context-aware metrics during translation generation, as detailed next.

MBR Decoding. Given a source text, x_t, the context, c_t, a set of candidate translations sampled from the model, $\mathcal{Y}_t$, and a context-aware metric, $\mathcal{M}$, the utility of each candidate $\hat{y}_t \in \mathcal{Y}_t$, is

$$u(\hat{y}_t) = \frac{1}{|\mathcal{Y}|} \sum_{y_t \in \mathcal{Y}_t} \mathcal{M}([c_t; x_t], [c_t; y_t], [c_t; \hat{y}_t]). \tag{2}$$

To determine the best translation, we then select the candidate that maximizes utility:

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

Note that the inference strategy can be employed independent of the training, i.e., by sampling from a non-context-aware distribution, $P(y_t|x_t)$ and using a context-aware metric for reranking. This approach can be beneficial when the context-aware metric captures complementary information not fully addressed during training, or when a context-aware MT model is unavailable.

Choice of Metric $\mathcal{M}$. Standard MT metrics often fall short in effectively utilizing context to determine translation accuracy (Voita et al., 2019). Thus, recent research has focused on designing metrics that better capture discourse information by using inter-sentential context (Vernikos et al., 2022; Jiang et al., 2022; Hu et al., 2023b; Fernandes et al., 2021). For example, context-aware extensions of metrics like COMET (Vernikos et al., 2022; Agrawal et al., 2024) DOCCOMET, or CONTEXTCOMET, compute quality scores for a source-reference-hypothesis tuple, $(x, y, \hat{y})$, using representations extracted from context-augmented inputs, $([c; x], [c; y], [c; \hat{y}])$, that correlate better with human judgments on document-level MT evaluation. We apply a similar approach by prepending the source from k previous turns, $x_{<t-k:t}$, to the tuple: $([x_{<t-k:t}; x_t], [x_{<t-k:t}; y_t], [x_{<t-k:t}; \hat{y}_t])$. Unlike Agrawal et al. (2024), who use hypotheses in the context, c_t, we use the original bilingual context to ensure that the added context remains independent of specific hypotheses from previous turns, enabling stable scoring in MBR decoding by leveraging shared discourse information.

4 Translation-Mediated Conversations: Case Studies

Translation-mediated conversations have diverse applications across numerous fields, including political, legal, medical, e-commerce, and everyday communication. In this work, we focus on two task-oriented applications, as detailed in § 4.1. We then present the evaluation setup and experimental settings in § 4.2 and § 4.3, respectively.

4.1 Application and Datasets

Customer-support Interaction. We use the dataset provided by the WMT 2024 Chat Shared Task (Mohammed et al., 2024), which includes real bilingual online customer service chats between an English-speaking agent and clients who speak Portuguese, French, Italian, Dutch, or Korean. The dataset spans several domains, including account registration issues, payment and delivery clarifications, and after-sale services in various industries, such as retail and gaming.

Personal-assistant Interaction. We use the BCONTRAST (Farajian et al., 2020) EN-DE dataset based on the Taskmaster-1 (Byrne et al., 2019) corpus. The dataset includes task-based bilingual dialogues in six domains: (i) ordering pizza, (ii) creating auto repair appointments, (iii) setting up ride service, (iv) ordering movie tickets, (v) ordering coffee drinks, and (vi) making restaurant reservations. This setup allows us to model structured human-assistant interactions, where the language model facilitates task completion in a controlled yet conversational manner. We note, however, that while we focus on task-oriented dialogue, our framework is designed to generalize to a wide range of LLM-driven interactions, including open-ended dialogue (e.g., with ChatGPT).

The general statistics from both datasets are presented 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

Dataset	Language Pair	# Instances			Avg. Source Length			Avg. # Segments per Conversation			% MuDA tagged	
		Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Dev	Test
WMT24	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
BConTrasT	en↔de	–	2100	–	–	43.03	–	–	26.92	–	–	22.86

Table 1: Statistics for each language pair and the data splits.

Model	Context?	en-xx			xx-en		
		chrF↑	Comet↑	MetricX↓	chrF↑	Comet↑	MetricX↓
Baselines							
GPT-4o	✗	70.09 7	92.62 5	0.37 5	77.33 5	92.41 4	0.50 3
	✓	70.34 7	92.93 4	0.33 3	74.75 8	91.59 6	0.58 4
TowerInstruct	✗	64.95 10	91.69 7	0.38 4	76.04 7	92.17 5	0.56 4
	✓	63.39 12	91.09 7	0.49 6	74.32 10	91.36 6	0.60 4
+ QAD (Comet)	✗	65.20 10	92.87 4	0.31 3	75.59 8	92.80 3	0.52 3
+ QAD (ContextComet)	✗	65.06 10	92.57 5	0.31 3	75.91 7	92.65 4	0.51 3
TowerChat	✗	71.68 7	93.01 5	0.32 3	77.97 5	92.72 4	0.51 3
	✓	75.93 4	93.63 3	0.32 3	78.87 3	93.01 3	0.47 2
	✓ (en)	74.81 4	93.33 3	0.33 3	78.76 3	92.90 4	0.47 2
+ QAD (Comet)	✓	76.36 2	**94.18** 1	**0.25** 2	**78.92** 2	**93.39** 2	**0.44** 1
+ QAD (ContextComet)	✓	**76.56** 2	94.05 2	0.26 2	**78.92** 2	93.24 3	**0.44** 1
+ SFT on QAD (ContextComet)	✗	73.04 6	93.25 4	0.31 3	78.40 4	92.81 4	0.50 2
	✓	76.22 3	93.72 3	0.29 3	78.80 3	93.04 3	0.47 2

Table 2: Main results on WMT24 Chat Shared Task. QAD with TowerChat significantly outperforms all baselines across the board. Models are grouped into statistically significant quality clusters. We bold-face both the best overall model and the best Tower-based model for each metric and language pair.

automatic tagger for identifying tokens belonging to certain discourse classes (lexical cohesion, verb forms, pronouns, formality) of potentially ambiguous translations (see Appendix B for more details). Tagging rules are validated by native speakers for linguistic accuracy. For each tag type, we report the F1 score based on matches between tagged words in the reference and hypothesis. While the WMT24 development and test sets exhibit a similar distribution regarding segment length and count, they differ significantly from the training dataset. Furthermore, up to 30% of en↔fr instances are flagged for disambiguation by MuDA, emphasizing the importance of context for generating high-quality translations.

4.2 Evaluation

Ambiguous contextual phenomena that require nuanced evaluation often arise in Chat MT. As such, we leverage three types of automatic evaluation: 1) for measuring overall translation quality, we use three metrics – two neural (Comet-22 by Rei et al. [2022], MetricX-XL by Juraska et al. [2023]) and one lexical (chrF by Popović [2015]); 2) a reference-free LLM-based metric based on GPT-4 that uses context for providing fine-grained error quality assessment following MQM typology (Agrawal et al., 2024, ContextMQM); 3) F1-score on MuDA tags for measuring whether models correctly resolve lexical ambiguities across diverse discourse phenomena.[3]

Considering all the metrics is crucial because Comet may favour our QAD strategies. On Tables 2 and 3, we report performance clusters based on statistically significant performance gaps at 95% confidence.[4] We create per-language groups for systems with similar performance, following

[3]We ignore conversational stopwords when measuring lexical cohesion: *um, uh, okay, ok, yes, no*.

[4]For segment-level metrics, such as Comet, we perform significance testing at the segment level. For chrF, we compute corpus-level scores calculated over 100 random samples, each with 50% of the total segments (without replacement).

Model	Context?	en-de			de-en		
		chrF↑	Comet↑	MetricX↓	chrF↑	Comet↑	MetricX↓
Baselines							
GPT-4o	✗	68.51 2	90.60 2	0.52 4	71.14 2	92.35 2	0.41 1
	✓	**70.23** 1	**90.96** 1	**0.39** 1	**72.72** 1	**92.81** 1	**0.39** 1
TowerInstruct	✗	62.46 13	88.20 6	0.54 4	69.54 6	91.98 4	0.47 4
	✓	62.75 12	88.29 6	0.56 5	**70.35** 3	91.99 4	0.46 3
+ QAD (Comet)	✗	63.42 11	90.15 3	0.46 2	69.24 8	92.31 2	0.44 2
+ QAD (ContextComet)	✗	64.10 10	89.78 4	0.46 2	69.39 7	92.22 3	0.44 2
TowerChat	✗	65.32 9	89.45 5	0.49 3	68.74 10	91.74 5	0.48 4
	✓	67.09 6	89.60 4	0.46 2	69.31 7	92.05 4	0.44 3
	✓ (en)	66.38 7	89.53 4	0.47 3	69.33 7	92.05 4	0.44 2
+ QAD (Comet)	✓	**67.84** 3	**90.94** 1	**0.40** 1	69.71 4	**92.34** 2	**0.43** 2
+ QAD (ContextComet)	✓	67.69 4	90.58 2	**0.40** 1	69.65 5	92.24 3	**0.43** 2
+ SFT on QAD (ContextComet)	✗	65.92 8	89.70 4	0.46 2	69.02 9	91.85 5	0.46 3
	✓	67.34 5	89.81 3	0.45 2	69.16 8	91.87 5	0.44 2

Table 3: Main Results on BConTrasT. QAD with TowerChat performs comparably with GPT-4o on Comet and MetricX. Models are grouped into statistically significant quality clusters. We bold-face both the best overall model and the best Tower-based model for each metric and language pair.

Freitag et al. (2023b), and obtain system-level rankings with the average of the obtained clusters, as Colombo et al. (2022). If no model wins on a majority of languages, there is no first cluster.

4.3 Experimental Settings

TowerChat. We finetune TowerBase 7B with TowerInstruct's hyperparameters (see Table 7) on the concatenation of TowerBlocks and the training set of the WMT24 shared task using context-aware prompts. Importantly, we do not use BConTrasT training data. This allows us to assess the model's generalization capabilities to a new domain. We report greedy and QAD results with the TowerChat-7B model. For QAD, we perform MBR with Comet or ContextComet on 100 candidates obtained via epsilon sampling with $\epsilon = 0.02$ (Hewitt et al., 2022). Epsilon sampling restricts token selection to those exceeding a probability threshold ϵ, reducing the risk of generating text that is too unreliable and was shown to be the more effective sampling strategy for MBR over alternatives by Freitag et al. (2023a).

Instruction Settings. To assess whether systems can properly leverage conversational context, we prompt the LLM-based MT with two instruction formats (see Figure 1): 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.[5] Additionally, we compare the latter to an alternative where the context is provided only in English: ✓**(en)**—retaining original texts in English and using translated texts for non-English source texts, as discussed in Section 3.1.

Baselines. We report greedy decoding with TowerInstruct-7B and GPT-4o.[6] The former serves as a direct baseline for our method, while the latter is a state-of-the-art baseline for MT (Sinitsyna and Savenkov, 2024). Furthermore, to assess whether QAD with context-aware metrics can improve translation quality without training, we also report QAD results with Comet or ContextComet for TowerInstruct.

Computational Complexity. MBR scales quadratically with the number of candidates, which can significantly slow down generation. This is particularly relevant in our application, where latency is a concern in real-world deployments. For example, in our setup, with 100 candidates, MBR requires, on average, 100 times more generation steps from the translation model and 10,000 forward passes

[5] During inference, the model generates {**target_seg**}.

[6] We used the snapshot `gpt-4o-2024-08-06`, with the same prompt as TowerInstruct without a chat template.

through the metric. However, the exact computation cost depends on hardware configuration and specific compute optimization. For example, we use the optimized implementation for MBR decoding with COMET that uses cached embedding representations. Recent work has introduced several techniques to improve the efficiency of MBR decoding, including distillation (Finkelstein and Freitag, 2024), low-rank factorization (Trabelsi et al., 2024), and linear-time approximations (Vamvas and Sennrich, 2024). While our primary focus is to show the effectiveness of context-aware QAD in improving translation quality and contextual accuracy, we include an additional experiment that applies training-time self-distillation on MBR outputs (Finkelstein and Freitag, 2024).

5 Main Results

Tables 2 and 3 present EN→XX and XX→EN results on WMT24 and BCONTRASTT, respectively.

TOWERCHAT Leverages Context more Adeptly than TOWERINSTRUCT. One of our goals was to create an LLM-based model that effectively leverages context to generate high-quality translations. As shown in Tables 2 and 3, TOWERCHAT outperforms TOWERINSTRUCT across both settings (*w/ context* and *w/o context*), language pairs, and evaluation metrics. The exception is the BCONTRASTT DE-EN setting, where TOWERINSTRUCT leads in CHRF, and lies in the same performance cluster as TOWERCHAT *w/ context* according to COMET and METRICX. Furthermore, TOWERCHAT shows an average improvement of 4 CHRF points for WMT24 EN-XX and 1.7 points EN-DE when using context (*w/ context*), compared to a context-agnostic prompt (*w/o context*). This trend also holds when evaluating translation quality with COMET, for 8 out of 10 WMT24 language pairs, as shown in the Appendix Table 10. We attribute this to the inclusion of context-augmented instruction dataset in TOWERCHAT's training, highlighting the effectiveness of in-domain fine-tuning. Moreover, incorporating bilingual context consistently improves translation quality over English-only context, yielding an average gain of 1.12 CHRF points for WMT24 EN-XX settings.

Leveraging Contextual Information is Particularly Helpful for Low-quality Translations. TOWERCHAT produces relatively better translations when provided with context than otherwise when

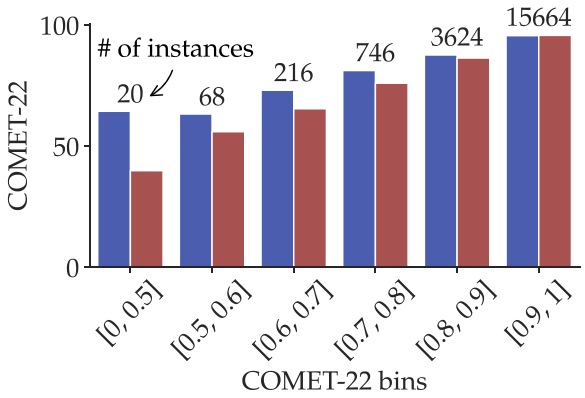


Figure 3: COMET from TOWERCHAT *w/o context*. Blue: w/ context. Red: w/o context.

the quality of the translation without context is low (see Figure 3), decreasing the likelihood of an unsuccessful interaction between participants. Gains are consistent across quality bins on both EN→XX and XX→EN language pairs (see Figure 11 in Appendix E). However, the impact of adding context diminishes as quality increases. This suggests that by developing quality estimation metrics for segment-level chat translation, one can design a dynamic selection mechanism that applies context only when the estimated quality without context falls below a certain threshold (Farinhas et al., 2025); we leave this to future work.

QAD Results in Consistent Gains over Greedy Decoding. In both datasets and for both Tower models, QAD consistently improves translation quality over greedy decoding across metrics. Furthermore, the highest-quality translations according to all metrics are obtained after performing QAD with COMET or CONTEXTCOMET on top of TOWERCHAT, even outperforming the GPT-4o baseline in the WMT24 dataset. Notably, the gains extend to METRICX, a metric not directly optimized by QAD, highlighting the robustness of the approach. In the BCONTRASTT EN-DE setting, QAD with COMET closes the gap with METRICX and COMET between TOWERCHAT (greedy) and GPT-4o models. This demonstrates how advanced inference methods can improve smaller models, enabling them to compete with larger models like GPT-4o.

MUDA-based Evaluation Validates our Findings. We report MUDA F1 scores between references and generated hypotheses for a

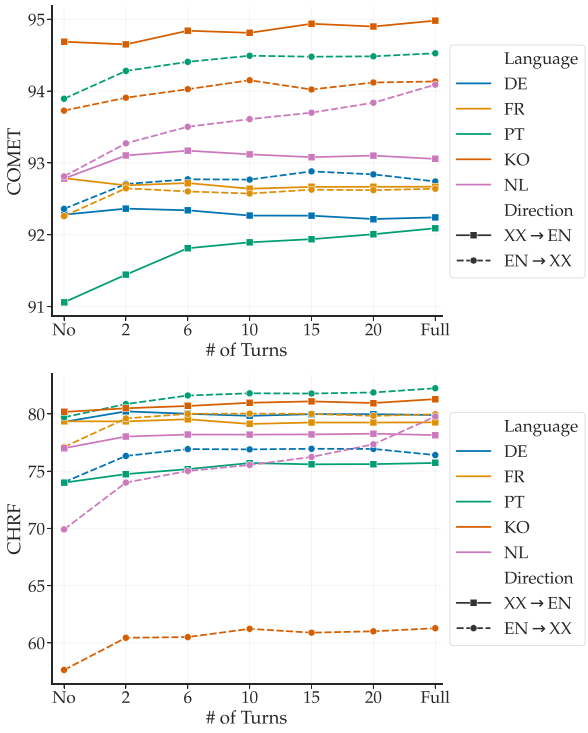


Figure 4: Comparison of chrF and Comet scores for varying context window sizes.

subset of models in Figure 4.[7] We can observe that, on average, across phenomena, on all datasets and language pairs: 1) TowerChat *w/ context* achieves higher F1 than TowerInstruct *w/o context*, confirming that TowerChat uses context to improve accuracy on discourse phenomena; 2) QAD-based approaches improve upon their respective greedy decoding counterparts. In 7 out of 12 settings (including QAD with both TowerChat and TowerInstruct), QAD with ContextComet (QAD-C) achieves higher F1 than QAD with Comet (QAD). This is also reflected in chrF scores per language pair (Appendix Table 9), where QAD-C outperforms QAD in 7 out of 10 settings.

Fine-grained Error Analysis Shows QAD-C Performs Better with Context-aware Hypotheses. To compare QAD and QAD-C, we obtain fine-grained MQM-like assessments with ContextMQM referred to as MQM (Table 4). Contrary to MuDA, MQM can identify errors that go beyond surface-level properties. Overall, leveraging QAD-C with TowerInstruct results in similar MQM scores as QAD on average; however, when the pool of candidates includes context-aware hypotheses as with TowerChat, QAD-C outperforms QAD on MQM evaluation in 9 out of 12 settings. To explain the effectiveness of QAD-C with TowerChat, we compare the quality of hypotheses generated by the two models as measured by chrF against references in Appendix § D: The candidates generated by TowerChat have a higher overlap with the reference than those generated by TowerInstruct. This shows that leveraging better context-aware metrics should further improve translation quality.

Distilling MBR Outputs Can Improve Quality while Reducing Computational Overhead. Having established the benefits of QAD with ContextComet in enhancing contextual accuracy—measured by improvements in MuDA and reductions in errors according to MQM—we now evaluate the impact of distilling this information via SFT using the WMT24 chat shared task dataset. Specifically, we fine-tune TowerChat for a second epoch, replacing the original chat data references with the model's own context-aware MBR outputs (SFT on QAD). This simple, single-pass distillation yields quality improvements over TowerChat across automatic metrics in out-of-English settings (see last two rows in Tables 2 and 3) while also enhancing MuDA accuracy (Table 4). These results prove that further efficiency gains in such systems are possible, though we leave deeper exploration of this to future work.

Context-augmented Instructions Drive Improvements Beyond In-domain Training. We conduct an additional ablation to confirm that the observed improvements are driven by the context-augmented instruction structure rather than by the exposure to in-domain data. In this setup, TowerInstruct is fine-tuned on a concatenation of TowerBlocks and the full WMT24 training dataset without any context-aware prompts (referred to as TowerInstruct (Chat)). As shown in Table 5, the context-aware TowerChat model consistently outperforms the non-contextual TowerInstruct (Chat) across both en-xx and xx-en directions, particularly for the latter. This is particularly evident when context-aware prompts are used during inference, where TowerChat achieves notable improvements in chrF and Comet scores and maintains competitive MetricX values. These results highlight that simply incorporating improved context-aware instructions during training,

[7] Phenomena-specific plots are presented in Figure 9.

Model	Context	WMT24 EN-DE			WMT24 EN-PT			WMT24 EN-FR			WMT24 EN-NL			WMT24 EN-KO			BConTrasT EN-DE		
		F1	XX	EN	F1	XX	EN	F1	XX	EN	F1	XX	EN	F1	XX	EN	F1	XX	EN
TowerInstruct	✗	78.29			76.56			80.44			52.40			52.44			69.47		
	✓	80.12			73.44			80.23			46.46			49.06			69.32		
+ QAD	✗	78.39	−0.210	**−0.458**	**77.64**	−0.841	−1.276	80.47	**−0.429**	−0.754	**54.80**	−0.384	−0.709	54.77	**−0.609**	**−0.862**	**71.63**	−0.574	−0.511
+ QAD-C	✗	**78.42**	**−0.173**	−0.463	76.47	**−0.776**	−1.298	**80.62**	−0.455	**−0.737**	52.57	**−0.375**	**−0.666**	**55.62**	−0.666	−0.870	69.76	**−0.559**	**−0.468**
TowerChat	✗	77.90			79.53			82.99			49.69			60.53			68.16		
	✓	79.81	−0.269	−0.416	86.55	−0.461	−0.755	86.34	−0.410	−0.635	68.27	−0.300	−0.452	60.12	−0.647	−0.852	**76.92**	−0.578	−0.476
+ QAD	✓	80.02	−0.176	−0.351	87.24	−0.441	−0.629	87.11	**−0.267**	−0.664	76.96	−0.236	**−0.457**	62.50	−0.349	**−0.874**	76.56	−0.417	−0.416
+ QAD-C	✓	80.68	**−0.153**	**−0.331**	85.94	**−0.438**	**−0.600**	87.17	−0.313	**−0.638**	**78.74**	**−0.219**	−0.481	**64.34**	**−0.318**	−0.948	74.50	**−0.397**	**−0.408**
+ SFT (QAD-C)	✓	**84.11**	−0.260	−0.407	**87.49**	−0.501	−0.608	**87.41**	−0.388	−0.635	71.07	−0.253	−0.503	63.94	−0.430	−0.906	76.12	−0.591	−0.461

Table 4: Context-based evaluation. QAD with ContextComet (QAD-C) outperforms QAD with Comet (QAD) on MuDA F1 and ContextMQM in 7/12 and 15/24 settings, respectively.

Context-aware?		en-xx			xx-en		
Training	Inference	chrF↑	Comet↑	MetricX↓	chrF↑	Comet↑	MetricX↓
✗	✗	75.80	93.54	**0.31**	77.80	92.80	0.50
✗	✓	72.89	92.73	0.38	75.39	91.21	0.59
✓	✗	71.68	93.01	0.32	77.97	92.72	0.51
✓	✓	**75.93**	**93.63**	0.32	**78.87**	**93.01**	**0.47**

Table 5: Ablation of using context-aware prompts during training and/or inference when TowerInstruct is finetuned with TowerBlocks and WMT24 Chat Datasets: TowerChat, trained with context-aware prompts, results in the best overall quality.

Models	en→xx	xx→en
TowerInstruct-7B	84.28	82.77
TowerChat-7B	83.95	82.54

Table 6: Comet scores for TowerInstruct and TowerChat on the WMT23 test set.

using the same in-domain dataset, enables the model to better focus on fine-grained details in the output, thereby enhancing its contextual understanding. Hence, it is not merely the quantity of domain-relevant data but how the model is guided to leverage contextual information that drives improvements.

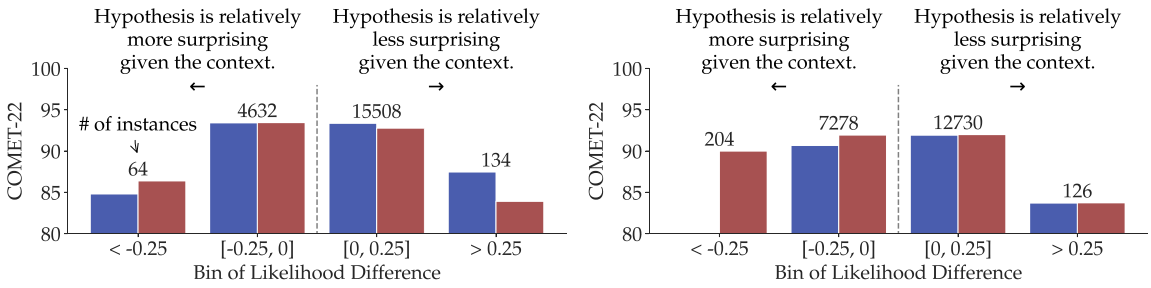
Finetuning on Chat Data Does Not Degrade General Translation Capabilities. To ensure that training on chat data did not impact translation capabilities on generic data, we report Comet on the standard WMT23 benchmark (Kocmi et al., 2023) averaged across en→xx and xx→en directions for TowerInstruct and TowerChat in Table 6. TowerChat suffers only minor degradation (-0.3) relative to TowerInstruct.

6 Assessing Context Usage

While our results show that using context in training and inference improves translation quality, it is still not clear when how context influences specific translations. As incorporating context in the translation process incurs additional computational costs, understanding when the context is used (beneficially) and what parts of this context are most influential could allow a more selective and efficient use of context. Additionally, validating that models use context in interpretable ways builds confidence in their reliability for real-world applications (Yin and Neubig, 2022; Briakou et al., 2023; Sarti et al., 2024; Cohen-Wang et al., 2024). Using the WMT24 Chat dataset, we analyze how much context is relevant for generating accurate and higher-quality translations, then assess whether the context is used meaningfully by TowerChat.

6.1 How Much Context is Needed?

Figure 4 shows the impact of varying the number of turns included in the context during inference on Comet and chrF. For both metrics, translation quality improves as the context window length increases. Notably, for en-nl and pt-en, using the full context yields improvements beyond those achieved with up to 20 turns of conversation. In contrast, incorporating 6 to 10 turns is sufficient for other language pairs to reach peak performance. Furthermore, we observe limited gains when adding context for generating translations into English, consistent with prior observations (Agrawal et al., 2024). These results suggest



(a) TOWERCHAT (b) TOWERINSTRUCT

Figure 5: COMET under different prompt settings (with and without context) for different bins of log LD on the hypothesis for TOWERCHAT (left) and TOWERINSTRUCT (right).

that the optimal context window varies by language pair and that adaptive strategies for optimal context selection may be useful.

6.2 How Does Context Influence Predictions?

To understand when context meaningfully influences translations, we employ measures that quantify the impact of context on model predictions:

- **P-CXMI** (Fernandes et al., 2023) measures how likely the *reference* (contextual) translation y is given a context, C, compared to when the context is not provided.

$$\text{P-CXMI} = \log \frac{P(y|x, C)}{P(y|x)} \quad (4)$$

  Intuitively, a higher P-CXMI means that the reference translation requires context to be translated accurately.

- **Likelihood Difference** (Shi et al., 2024) We measure the difference between the log-likelihood of a context-aware hypothesis against a context-agnostic hypothesis as:

$$\log \text{LD} = \log \frac{P(\hat{y}_{\text{ctx}}|x, C)}{P(\hat{y}_{\text{no-ctx}}|x)} \quad (5)$$

  Unlike P-CXMI, this metric does not rely on a reference translation. Instead, it directly evaluates how incorporating context affects the likelihood of the generated hypothesis.

Our findings on how these metrics relate to translation quality are presented below:

Quality is Higher for Sentences that Require Resolving Ambiguity by P-CXMI. Figure 6 shows that TOWERCHAT performs better with

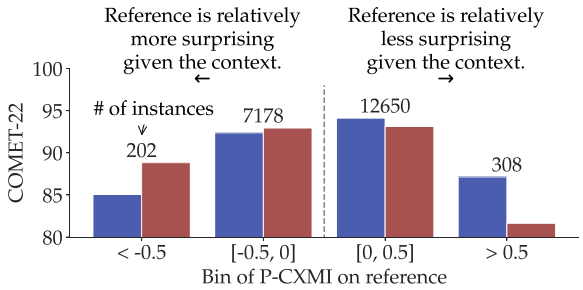


Figure 6: COMET under different prompts (with and without context) for different P-CXMI bins.

context when P-CXMI is positive and worse otherwise. In other words, hypotheses are better on average when a reference translation requires context according to the model. The higher/lower the P-CXMI, the more positive/negative the change in quality.

The Output Likelihoods of TOWERCHAT Predict the Impact of Context on Quality. Figure 5a shows that translation quality is higher for the context-aware hypothesis when it is more likely than the context-unaware hypothesis and vice-versa. This means that we can predict, to a certain extent, whether a translation will benefit from context when using TOWERCHAT. Remarkably, this does not hold for TOWERINSTRUCT (Figure 5b): Regardless of likelihoods, translation quality is always highest when not leveraging context.[8]

[8]Except for a small difference in the last bin of likelihood difference, this is also the case for the version of TOWERINSTRUCT trained on context-unaware chat data (see Figure 10).

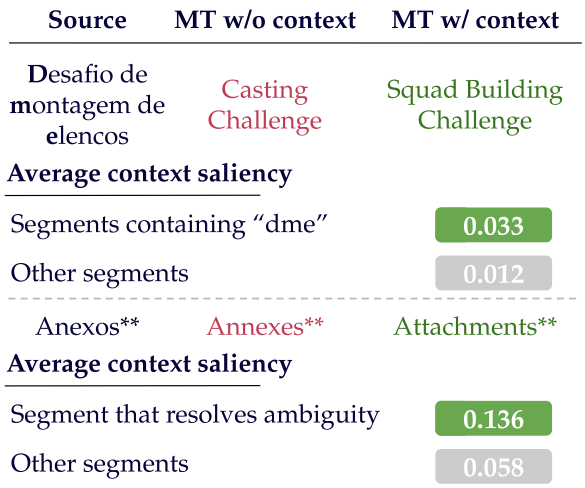


Figure 7: Two examples of PT→EN translations with and without context where the contextually informed translation is accurate, while the translation without context is lexically correct but contextually incorrect: saliency values are high for context segments that resolve semantic ambiguity.

6.3 Which Parts of the Context Impact Target Predictions?

PeCoRE (Sarti et al., 2024) is a method to identify salient context input tokens that explain the generation of context-sensitive output tokens over alternatives. Following their notation, we use $X_{ctx} : \{X, C\}$ and $X_{no-ctx} : X$ to represent the context-aware and context-agnostic inputs. These inputs are passed through TowerChat to generate context-sensitive $\hat{y}_{ctx}$ and context-agnostic $\hat{y}_{no-ctx}$ outputs respectively. At each decoding step i, Let P^i_{ctx} represent the token-level probability distribution obtained from the model when conditioned on $\{X_{ctx}, \hat{y}_{ctx,<i}\}$. To form a contrastive target distribution, we force-decode the non-contextual output, y_{no-ctx} with the context-aware input, X_{ctx}, resulting in the probability distribution, $P^i_{\text{no-ctx}}$. The degree to which individual context tokens contribute to the contextualized generation is quantified by taking the gradient of a target function (f_{tgt}) with respect to each input token. We use the likelihood ratio between P^i_{ctx} and $P^i_{\text{no-ctx}}$ as f_{tgt}.

Salient Tokens Signal Important Parts of Context Leveraged by TowerChat. In Figure 7 we give two examples of translations produced by TowerChat with and without context. The former is correct, while the latter is lexically correct but semantically incorrect. We measure the saliency of segments in the context crucial to resolving the source text's semantical ambiguity. TowerChat shows high—almost threefold—saliency for these segments compared to the rest. In the first case, ''dme'' is an acronym for the source text. It co-occurs with segments that make it clear that the conversation is about squads and not casts; TowerChat with context is able to pick up on this and produce a correct translation. In the second case, a segment in the context makes it obvious that ''Anexos'' should be translated to ''Attachments'' (as in email) rather than ''Annexes''. This segment is much more salient than the rest. This shows that the model can correctly use the context it is provided with to generate correct translations in both lexical and contextual senses. We provide some additional examples along with the complete conversation context for the ones presented here in Appendix F.

7 Conclusion and Future Work

This work presents a context-aware framework for improving translation quality in bilingual conversations. Experiments on two task-oriented domains show that the resulting model is better at leveraging contextual information during training and inference. Quality-aware decoding methods with hypotheses generated by a context-aware model further improve translation quality and accuracy in modeling discourse phenomena over strong baselines for both domains and all language pairs.

However, several challenges remain. Our results indicate that context benefits low-quality segments most, and adding context beyond a specific number of turns yields limited gains. This suggests that context should be selectively incorporated where it has the most significant impact. One can design a dynamic selection mechanism by being strategic about the number of turns included and/or using more effective quality estimation metrics for segment-level chat translation—particularly those aligned with reference-based metrics like COMET-22. Such a mechanism would apply context only when the estimated quality of a segment falls below a certain threshold (Farinhas et al., 2025), ensuring more efficient and targeted use of contextual information. We could also use such a metric to perform QE reranking, which scales linearly with the number of candidates instead of MBR decoding. This would optimize computational efficiency while

preserving translation quality where it matters most. We leave a detailed investigation of these strategies to future work.

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267. https://doi.org/10.1162/tacl_a_00700

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.564

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. https://doi.org/10.18653/v1/2023.findings-emnlp.744

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. *arXiv preprint, arXiv:2402.17733*.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with contrastive phrasal highlighting: A case study in assisting humans to detect translation differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.690

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1459

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages 1–31. https://doi.org/10.1017/nlp.2024.7

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://doi.org/10.52202/079017-3035

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*.

Donald A. DePalma, Benjamin B. Sargent, and Renato S. Beninatto. 2006. Can't read, won't buy: Why language matters on global websites. *Lowell, MA: Common Sense Advisory Inc*.

Krishno Dey, Prerona Tarannum, Md. Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in English: Evaluation of large language models on English, low-resource and cross-lingual settings. *arXiv preprint, arXiv: 2410.13153*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*. https://doi.org/10.18653/v1/2024.naacl-short.46

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-short.46

M. Amin Farajian, António V. Lopes, André F. T. 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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.wmt-1.3

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. https://doi.org/10.18653/v1/2023.emnlp-main.733

António Farinhas, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, and André F. T. Martins. 2025. Translate smart, not hard: Cascaded translation systems with quality-aware deferral. *arXiv preprint arXiv:2502.12701*. https://doi.org/10.18653/v1/2025.emnlp-main.1358

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. https://doi.org/10.18653/v1/2022.naacl-main.100

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. 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

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023a. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.617

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. https://doi.org/10.1162/tacl_a_00491

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

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

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. https://doi.org/10.1162/tacl_a_00642

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. https://doi.org/10.18653/v1/2022.findings-emnlp.249

Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023a. A systematic study of performance disparities in multilingual task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.422

Xinyu Hu, Xunjian Yin, and Xiaojun Wan. 2023b. Exploring context-aware evaluation metrics for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15291–15298, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1021

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

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, pages 2627–2638. https://doi.org/10.1145/3589334.3645643

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

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. https://doi.org/10.18653/v1/2023.wmt-1.41

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

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. https://doi.org/10.18653/v1/2023.wmt-1.1

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich,

Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Preliminary WMT24 ranking of general MT systems and llms. *arXiv preprint arXiv:2407.19884*.

Lea Krause, Selene Baez Santamaria, and Jan-Christoph Kalo. 2024. Graph representations for machine translation in dialogue settings. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1038–1046, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.106

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat 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 5711–5724, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.444

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? A study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, 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, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2). https://doi.org/10.1145/3441691

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.712

John Mendonca, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C. Farinha, Helena Moniz, Alon Lavie, and Isabel Trancoso. 2023. Dialogue quality and emotion annotations for customer support conversations. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 9–21, Singapore. Association for Computational Linguistics.

Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C. Farinha, and José G. C. De Souza. 2024. Findings of the WMT 2024 shared task on chat translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 701–714, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.59

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.82

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.653

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. https://doi.org/10.18653/v1/2022.wmt-1.26

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–483, Nusa Dua, Bali. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.31

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

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. https://doi.org/10.18653/v1/2022.wmt-1.52

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. *arXiv preprint, arXiv:2309.08590*.

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*, FAccT '22, page 2223–2238, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3531146.3534638

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-short.69

Daria Sinitsyna and Konstantin Savenkov. 2024. Comparative evaluation of large language models for linguistic quality assessment in machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 154–183, Chicago, USA. Association for Machine Translation in the Americas.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2058

Mingi Sung, Seungmin Lee, Jiwon Kim, and Sejoon Kim. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1011–1015, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.102

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. 2024. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pages 54714–54733. Curran Associates, Inc. https://doi.org/10.52202/079017-1734

Marcos Treviso, Nuno M. Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André F. T. Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. *arXiv preprint, arXiv: 2406.19482*. https://doi.org/10.18653/v1/2024.findings-emnlp.892

Jun-ichi Tsujii and Makoto Nagao. 1988. Dialogue translation vs. text translation. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*. https://doi.org/10.3115/991719.991778

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.71

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. https://doi.org/10.18653/v1/2022.wmt-1.6

Sebastian T. Vincent, Loïc Barrault, and Carolina Scarton. 2022. Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 121–130, Ghent, Belgium. European Association for Machine Translation.

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

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. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1301

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2748–2754, Portorož, Slovenia. European Language Resources Association (ELRA).

Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

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.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024.

Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.14

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Bryan Zhang and Amita Misra. 2022. Machine translation impact in E-commerce multilingual search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 99–109, Abu Dhabi, UAE. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-industry.8

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

Yongle Zhang, Dennis Asamoah Owusu, Marine Carpuat, and Ge Gao. 2022. Facilitating global team meetings between language-based subgroups: When and how can machine translation help? *Proceedings of ACM on Human-Computer Interaction*, 6(CSCW1). https://doi.org/10.1145/3512937

Yongle Zhang, Dennis Asamoah Owusu, Emily Gong, Shaan Chopra, Marine Carpuat, and Ge Gao. 2021. Leveraging machine translation to support distributed teamwork between language-based subgroups: The effects of automated keyword tagging. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3411763.3451837

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

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. https://doi.org/10.18653/v1/2024.findings-naacl.176

A Context Prompt Example

Context: Naja es geht so.
Ich habe gestern einen ärgerlichen Vorfall.
Ich hatte auf meinem ACC knapp 335000 PRS-ORG Coins und beim anmelden hatte Ich nur noch 776
So you're missing your coins.
That's indeed concerning.
And I'll surely look into this.
Please provide me the email of the account.
Thank you.
Let me check if there were any transaction for coins that were not done by you.
Thank you.
I can see there are no suspicious activity on your account in past 7 days.
I can see all the coins were used by your web app and PRS-ORG.

Translate the English source text to German, given the context.
English: Let me tell you where it was used.
German: Lassen Sie mich Ihnen sagen, wo es verwendet wurde.

Figure 8: Specific training instance with context for TowerChat. Gradient updates are only performed on the reference, and new lines are encoded as \n.

B MuDA-based Evaluation

MuDA uses a series of rule-based taggers to identify tokens in reference and hypothesis translations that are related to specific discourse phenomena. These tokens often correspond to tokens where discourse coherence is particularly challenging, such as pronouns (e.g., ''it'' in English, which could translate to ''er,'' ''sie,'' or ''es'' in German depending on the antecedent's gender). Native speakers verify the tagging rules to ensure linguistic validity. We then compute the F1 score for each tag based on whether a tagged word in the reference/hypothesis also appears and is tagged in the hypothesis/reference.

C TowerChat Training Hyperparameters

We do full finetuning of TowerBase with the hyperparameters of TowerInstruct (Alves et al., 2024) (see Table 7).

Precision	bfloat16
Packing	True
Global train batch size	256
Number of Epochs	4
Learning rate	7e-6
LR Scheduler	cosine
Warmup Steps	500
Weight Decay	0.01
Optimizer	Adam (Kingma and Ba, 2015)
Adam (β_1, β_2, ϵ)	(0.9, 0.999, 1e-8)
Maximum Sequence Length	2048

Table 7: Hyperparameter configuration.

D Oracle Analysis - chrF

We present an oracle analysis on the quality of hypotheses generated by the TowerChat and TowerInstruct as measured by chrF in Table 8: The candidates generated by TowerChat have a higher overlap with the reference than those generated by TowerInstruct.

LP	EN-XX			XX-EN		
	TowerChat	TowerInstruct	Δ	TowerChat	TowerInstruct	Δ
DE	90.21	88.92	+1.29	91.77	92.92	−1.15
FR	90.65	89.89	+0.76	93.19	94.91	−1.72
KO	83.18	67.15	+16.03	93.56	92.15	+1.41
NL	91.82	83.69	+8.13	91.93	91.25	+0.68
PT	94.04	87.03	+7.01	91.99	86.30	+5.69

Table 8: Oracle chrF scores on the pool of candidates generated by the two configurations: TowerChat with context-aware prompt and TowerInstruct with context-agnostic prompts, respectively.

E Test Results and Analysis by Language Pair

Tables 9, 10, and 11 show chrF, Comet, and MetricX scores for individual language pairs across evaluated settings.

Model	Context?	EN-XX					XX-EN				
		DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines											
GPT-4o	✗	76.67	78.69	75.81	47.63	71.65	79.60	78.27	73.73	76.45	78.62
	✓	74.96	78.45	76.05	48.78	73.49	78.01	77.64	72.58	69.21	76.28
TowerInstruct	✗	71.81	74.59	72.26	43.18	62.90	77.57	79.02	72.06	75.73	75.80
	✓	71.16	74.38	68.50	41.70	61.23	75.68	78.31	71.83	72.63	73.15
+ QAD (Comet)	✗	72.74	73.60	72.95	43.11	63.59	76.26	79.44	71.86	75.32	75.07
+ QAD (ContextComet)	✗	72.05	74.08	72.96	43.19	63.02	76.80	79.13	72.15	75.88	75.61
TowerChat	✗	74.04	77.12	79.71	57.63	69.91	79.31	79.36	74.00	80.17	77.01
	✓	76.41	79.97	82.24	61.28	79.78	79.91	79.26	75.72	81.30	78.15
	✓ (en)	77.31	80.27	81.67	61.06	73.73	79.87	79.14	74.97	81.52	78.31
+ 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
+ SFT on QAD (ContextComet)	✗	72.30	77.25	79.74	58.46	69.74	78.52	79.78	73.66	80.49	77.17
	✓	76.36	79.67	81.82	61.09	78.62	79.01	78.52	75.64	80.52	77.33

Table 9: Results by chrF (↑) on WMT24 Chat Shared Task dataset by language pair.

Model	Context?	EN-XX					XX-EN				
		DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines											
GPT-4o	✗	92.74	92.43	93.01	92.26	92.68	92.16	92.18	91.40	93.24	93.07
	✓	92.49	92.62	93.40	93.06	93.08	91.95	91.76	91.19	90.67	92.39
TowerInstruct	✗	91.71	91.89	91.90	91.64	91.30	92.08	92.78	90.43	93.13	92.45
	✓	91.48	91.08	90.79	91.13	91.00	91.33	91.89	90.63	91.88	91.08
	✓ (en)	92.94	92.70	94.38	94.10	92.51	92.30	92.68	91.67	94.87	92.98
+ QAD (Comet)	✗	92.77	92.62	93.24	93.04	92.68	**92.78**	**93.34**	91.13	93.87	92.88
+ QAD (ContextComet)	✗	92.53	92.49	92.86	92.75	92.23	92.58	93.2	90.87	93.81	92.77
TowerChat	✗	92.36	92.26	93.89	93.73	92.81	92.28	92.79	91.06	94.69	92.78
	✓	92.74	92.64	94.53	94.13	94.09	92.24	92.67	92.09	94.98	93.06
+ QAD (Comet)	✓	**93.28**	**93.13**	**94.91**	**95.01**	**94.54**	92.58	92.95	**92.63**	**95.32**	**93.49**
+ QAD (ContextComet)	✓	93.22	92.96	94.76	94.96	94.36	92.48	92.71	92.46	95.16	93.38
+ SFT on QAD (ContextComet)	✗	92.55	92.68	93.96	93.96	93.10	92.34	92.93	91.18	94.71	92.87
	✓	92.96	92.84	94.39	94.40	94.02	92.27	92.69	92.13	95.05	93.04

Table 10: Results by Comet (↑) on WMT24 Chat Shared Task dataset by language pair.

Model	Context?	en-xx					xx-en				
		DE	FR	PT	KO	NL	DE	FR	PT	KO	NL
Baselines											
GPT-4o	✗	0.42	0.28	0.29	0.51	0.33	0.52	0.51	0.67	0.32	0.46
	✓	0.45	0.22	0.28	0.39	0.30	0.55	0.56	0.64	0.65	0.50
TowerInstruct	✗	0.28	0.23	0.43	0.57	0.37	0.50	0.53	0.86	0.37	0.53
	✓	0.38	0.29	0.69	0.60	0.49	0.56	0.55	0.74	0.46	0.69
+ QAD (Comet)	✗	**0.25**	0.20	0.37	0.42	0.31	0.48	0.49	0.78	0.35	0.50
+ QAD (ContextComet)	✗	0.26	**0.19**	0.33	0.43	0.31	0.47	0.48	0.79	0.33	0.50
TowerChat	✗	0.27	0.24	0.29	0.42	0.37	0.50	0.51	0.71	0.33	0.52
	✓	0.34	0.26	0.27	0.45	0.27	0.47	0.48	0.60	0.30	0.48
	✓ (en)	0.26	0.24	0.28	0.46	0.41	0.49	0.47	0.60	0.30	0.49
+ 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**
+ SFT on QAD (ContextComet)	✗	0.29	0.23	0.26	0.38	0.37	0.50	0.50	0.73	0.33	0.52
	✓	0.31	0.23	0.29	0.37	0.26	0.50	0.49	0.59	0.32	0.51

Table 11: Results by MetricX (↓) on WMT24 Chat Shared Task dataset by language pair.

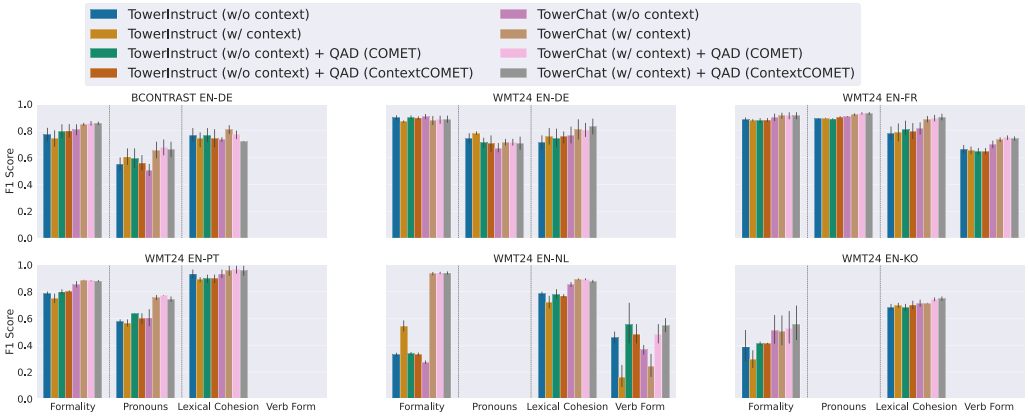


Figure 9: MuDA F1 by language pairs. LC: Lexical Cohesion, VF: Verb Form, P: Pronouns, F: Formality. On average, QAD with ContextComet has the best F1 score in 7 out of 12 settings.

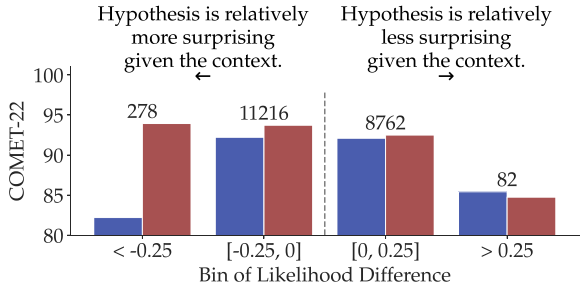


Figure 10: Comet under different prompts (with and without context) for different P-CXMI bins for TowerInstruct trained on chat domain data but context-unaware.

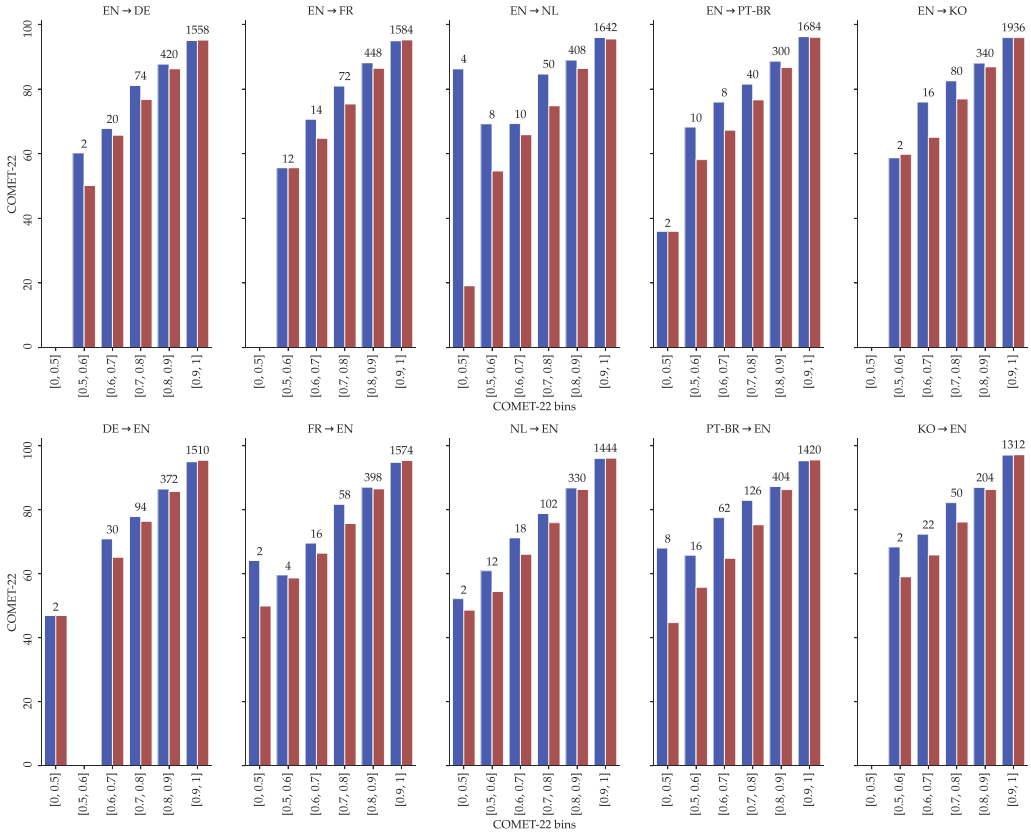


Figure 11: Quality Bins from COMET for TOWERCHAT *w/o context* on EN→XX (top) and XX→EN (bottom) language pairs. Blue: w/ context. Red: w/o context.

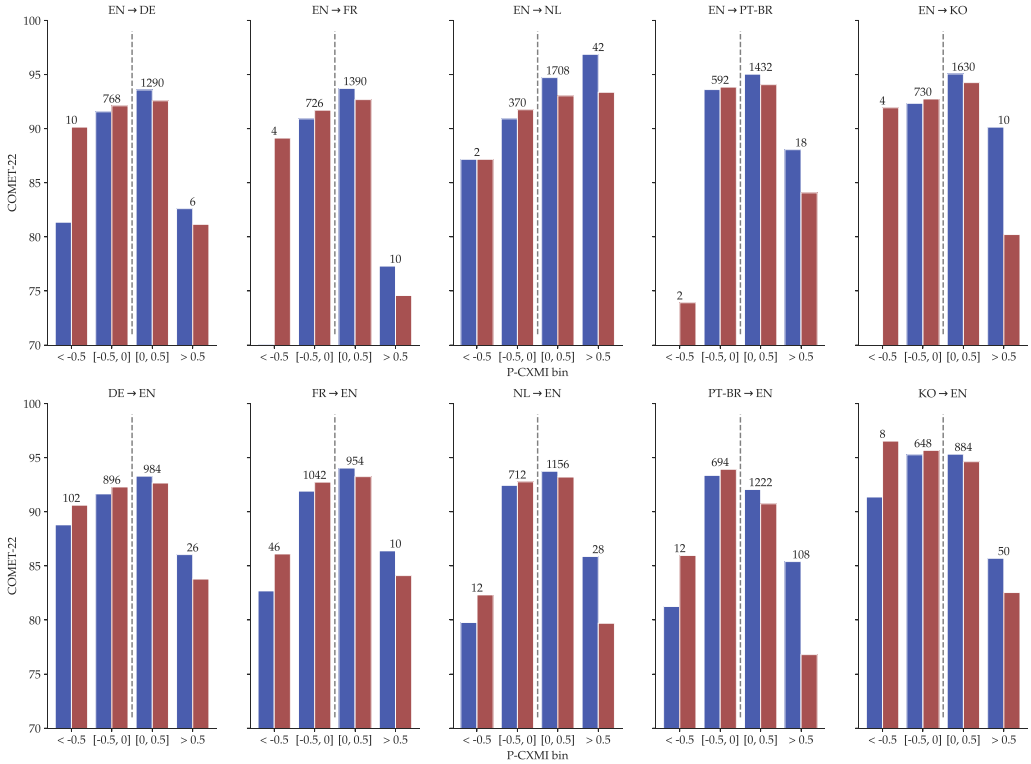


Figure 12: COMET under different prompt settings (with and without context) for different P-CXMI bins and language pairs.

F Additional Examples

We show concrete examples where context results in a more contextually appropriate translation in Figures 13, 14, 15, and 16.

Context: Obrigado por entrar em contato com a Ajuda PRS-ORG, o meu nome é NAME-N.
Para começar, pode-me indicar o seu nome por favor?
NAME-M
Hello NAME-M!
Its pleasure chat with you, how are you doing today?
Olá, tudo bem?
Gostaria de ajuda sobre o dme do PRS-ORG
Eu comecei a fazer o dme mas ele pede cartas que estão fora de packs e extintas no mercado
You are an incredible human!
Thank you so much for asking.
I am doing well and everything would be better if more people were like you.
Eu ja fiz 13 partes do dme e agora não consigo fazer o resto
Não acho justo com os jogadores manter um dme tão caro que seja impossível de fazer
Gostaria de receber os jogadores de volta ou receber os que faltam, no caso capitaes do PRS-ORG
Please do not worry, and be rest assured, I will try my best to resolve your issue regarding PRS-ORG
and cheer you up!
May I know what you referring with PRS-ORG?
I am sorry the term is not clear to me.

Translate the Brazilian Portuguese source text to English, given the context.
Brazilian Portuguese: Desafio de montagem de elencos
English:

TOWERCHAT w/o Context: Casting Challenge
TOWERCHAT w/ Context: Squad Building Challenge

Figure 13: Full conversation of first example in Figure 4. In red, the translation without context, which is wrong; in green, the translation with context, which is correct.

Context: Obrigado por entrar em contato com a Ajuda PRS-ORG, o meu nome é NAME-M.
Para começar, pode-me indicar o seu nome por favor?
Olá NAME-M!
Meu nome é NAME-M.
Hey NAME-M, nice meeting you.
Hope you are doing good today.
Sim estou lhe desejo que esteja bem também NAME-M. Consegues ver os dos que enviei?

Translate the Brazilian Portuguese source text to English, given the context.
Brazilian Portuguese: Anexos**
English:

TOWERCHAT w/o Context: Anexos**
TOWERCHAT w/ Context: Attachments**

Figure 14: Full conversation of second example in Figure 4. In red, the translation without context, which is wrong; in green, the translation with context, which is correct.

Context: Bom dia Obrigado por entrar em contato com a Ajuda PRS-ORG, o meu nome é NAME-M.
Para começar, pode-me indicar o seu nome por favor?
Meus PRS-ORGs points que comprei não caiu
Efetuei o pagamento e foi descontado
Mais ainda não caiu no jogo
Hello, nice to meet you.
I will surely try my best to help you with missing PRS-ORG Mobile points.
May I know when the purchase was made?
Então?
Hoje
Thanks for the details.
There's nothing to be worried about.

Translate the Brazilian Portuguese source text to English, given the context.
Brazilian Portuguese: Tou vendo
English:

TOWERCHAT w/o Context: I'm watching.
TOWERCHAT w/ Context: I see

Figure 15: Full conversation where context helps improve the translation quality. In red, the translation without context, which is wrong; in green, the translation with context, which is correct. In the translation without context, ''vendo'' is translated to ''watching''; this could be correct in certain contexts, but in this particular one—where the customer simply wants to acknowledge what the agent said—it is not. Instead, ''I see'' is the correct translation.

Context: Obrigado por entrar em contato com a Ajuda PRS-ORG, o meu nome é NAME-M.
Para começar, pode-me indicar o seu nome por favor?
NAME-F
não consigo realizar o pagamento dos pacotes, dá não autorizado sendo que possuo crédito no cartão
Hello NAME-F, hope you are fine, how may I help you?
Hello NAME-F, hope you are fine.
I see, let me check if we can fix the issue that you are facing with making purchase.
Please share the email address linked to the account.
EMAIL
So what exactly happens when you go to make transaction?
passa o cartão, normalmente
segundos depois
a compra é estornada
tentei 2 cartões diferentes
Let me check.
On my end it comes as the transaction is pending.
o que devo fazer?
tentar depois?
Yes.

Translate the English source text to Brazilian Portuguese, given the context.
English: Give it 24 hours cool down time.
Brazilian Portuguese:

TOWERCHAT w/o Context: Deixe por 24 horas para esfriar.
TOWERCHAT w/ Context: Dê um período de 24 horas de espera.

Figure 16: Full conversation where context helps improve the translation quality. In red, the translation without context, which is wrong; in green, the translation with context, which is correct. In the translation without context, ‘‘cool down’’ is literally translated to ‘‘esfriar’’, which is incorrect in context—the agent is telling the customer to wait (‘‘espera’’).