# Findings of the WMT 2024 Shared Task on Chat Translation

**Wafaa Mohammed**[1] **Sweta Agrawal**[2] **M. Amin Farajian**[3]
**Vera Cabarrão**[3] **Bryan Eikema**[1] **Ana C. Farinha**[3] **José G. C. de Souza**[3]

[1]University of Amsterdam, Netherlands
[2]Instituto de Telecomunicações, Lisbon, Portugal
[3]Unbabel, Lisbon, Portugal

## Abstract

This paper presents the findings from the third edition of the Chat Translation Shared Task. As with previous editions, the task involved translating bilingual customer support conversations, specifically focusing on the impact of conversation context in translation quality and evaluation. We also include two new language pairs: English↔Korean and English↔Dutch, in addition to the set of language pairs from previous editions: English↔German, English↔French, and English↔Brazilian Portuguese.

We received 22 primary submissions and 32 contrastive submissions from eight teams, with each language pair having participation from at least three teams. We evaluated the systems comprehensively using both automatic metrics and human judgments via a direct assessment framework. The official rankings for each language pair were determined based on human evaluation scores, considering performance in both translation directions—agent and customer. Our analysis shows that while the systems excelled at translating individual turns, there is room for improvement in overall conversation-level translation quality.

## 1 Introduction

Translating conversational text, in particular customer support chats, is an important and challenging application for machine translation (MT) technology. According to a 2020 survey from CSA Research, 75% of shoppers are more likely to make another purchase if customer support is offered in their native language, making it appealing for businesses to invest in multilingual support.[1] However, there are several key challenges to translating chats: customer support chats typically feature short text exchanges between agents and customers (see Table 1), leading to fragmented sentences and omission of information (implied by the context). This makes it difficult for MT systems to produce coherent translations that maintain the intended meaning of the text (Farajian et al., 2020). Furthermore, chats often use colloquial language and are characterized by informality and grammatical inaccuracies (Gonçalves et al., 2022). Consequently, translating such content poses a dual challenge: not only must a system accurately translate between languages, but it should also effectively model the nuances and ambiguity in a dialogue.

While recent advancements in MT systems, driven by LLMs, have proven effective in various tasks, bilingual chat translation remains underexplored. The **Chat Translation Shared Task** aims to bridge this gap by promoting research and development of MT systems designed specifically for conversational translation. This year's edition places special emphasis on the role of conversation context, encouraging teams to examine how context influences translation in the inherently ambiguous and dynamic nature of chat interactions. Following the success of the previous two editions of the Chat Translation Shared Task (Farajian et al., 2020; Farinha et al., 2022), this year we organized the third edition of the task with the following improvements:

- We expanded the set of language pairs to include English↔Korean (EN-KO) and English↔Dutch (EN-NL), in addition to languages from previous editions: English↔German (EN-DE), English↔French (EN-FR), and English↔Brazilian Portuguese (EN-PT).
- We carefully curated the evaluation sets to enable the evaluation of effective context utilization on systems' performance.
- We conducted a comprehensive evaluation of all systems using: a) automatic metrics (both neural and lexical) that assess translation quality and the accuracy of modeling discourse phenomena

[1] https://csa-research.com/Featured-Content/For-Global-Enterprises/Global-Growth/CRWB-Series/CRWB-B2C



*Proceedings of the Ninth Conference on Machine Translation*, pages 701–714
November 15-16, 2024 ©2024 Association for Computational Linguistics

| | |
|---|---|
| customer | Hallo, ich komme nicht in meine Sum up pos was denn no App rein<br>Hello, I can not get into my sum up pos what then no app |
| agent | I am sorry to hear that.<br>Es tut mir leid, das zu erfahren. |
| agent | Let me see what I can do for you<br>Lassen Sie mich sehen, was ich für Sie tun kann. |
| agent | Could you please tell me what error message you can see while logging in to your POS?<br>Könnten Sie mir bitte sagen, welche Fehlermeldung Sie sehen können, während Sie sich bei Ihrem POS anmelden? |
| customer | Wenn ich auf die App gehe, erscheint dieses Gerät hinzufügen.<br>When I go to the app, it shows Add this device. |
| agent | Could you please try to connect the App with the POS?<br>Könnten Sie bitte versuchen, die App mit dem POS zu verbinden? |
| customer | die App ist die PRS-ORG pos app<br>the app is the PRS-ORG app |
| customer | ich habe die Frage daher nicht verstanden<br>so I did not understand the question |
| agent | Could you please elaborate on your query?<br>Könnten Sie bitte Ihre Anfrage näher erläutern? |

Table 1: An example of a EN-DE conversation between a *customer* (👤) and an *agent* (🎧) from MAIA dataset.

using MUDA (Fernandes et al., 2023b), b) human direct assessments by professional linguists, and c) LLM-based fine-grained error analysis following the MQM framework.

We received a total of 22 primary submissions, 6 submissions for en↔de, 5 for en↔fr, 4 for en↔nl, 4 for en↔pt-br, and 3 for en↔ko. Six out of the eight teams used large language models (LLMs) as their base translation model, implementing various strategies such as finetuning on shared task data, augmenting training data with synthetic datasets, prompting strategies, quality-aware decoding, and several ways of leveraging conversational context to improve translation quality. With these multi-faceted solutions explored by several teams, this year's shared task yields valuable insights into the effectiveness of LLMs in translating conversational texts. We summarize the key findings from the shared task below:

- Incorporating contextual information from previous turns almost always improved translation quality. However, the optimal method for introducing context (whether through summary, graph, or raw context) still requires further investigation.
- Human evaluation showed that turn-level translation quality was consistently high across all participating systems and language pairs. Nonetheless, there is room for improvement in translating texts from later turns and at the conversation level as a whole.
- The UNBABEL-IT submission achieved the best results across most language pairs and evaluation criteria, except on the EN-DE and EN-FR tasks according to automatic metrics.

These findings suggest that future editions of the shared task could benefit from a) designing evaluation frameworks, both automatic and human, that specifically target dialogue-specific criteria to better understand system limitations (Yeh et al., 2021; A, 2022; Deriu et al., 2021); b) expanding the datasets to include more challenging domains (e.g. patient-physician conversation or everyday dialogues) and contexts (e.g. multimodal chats) for a more thorough evaluation of MT systems.

## 2 Task Description

As in previous editions of the task, we evaluate the effectiveness of a translation layer in translating text from the customer's language to the agent's language (e.g., English) and vice versa. We provide real bilingual customer support data for five different language pairs and encourage the participants to use conversation context. They are asked to submit translations for both directions (agent and customer). We detail the shared task dataset provided to the participants and evaluation in § 2.1 and § 2.2 respectively.

### 2.1 Data: The MAIA 2.0 Corpus

The MAIA 2.0 corpus builds upon the dataset released in the previous edition (Farinha et al., 2022) and includes two additional language pairs: Dutch and Korean. Furthermore, we expanded the sizes of the existing language pairs, ensuring that each language pair contained approximately 20k segments. The dataset encompasses dialogues across diverse

| LP | train | | | | dev | | | | test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # seg | # conv | # length | # words | # seg | # conv | # length | # words | # seg | # conv | # length | #words |
| EN-NL | 15.5k | 595 | 26.0 | 8.6 | 2.5k | 72 | 35.4 | 9.8 | 2k | 58 | 34.7 | 10.2 |
| EN-PT | 15.0k | 435 | 34.7 | 8.0 | 2.5k | 96 | 26.6 | 8.8 | 2k | 73 | 27.9 | 8.8 |
| EN-DE | 17.8k | 493 | 36.1 | 8.5 | 2.5k | 82 | 31.3 | 9.4 | 2k | 67 | 30.5 | 9.4 |
| EN-KO | 16.1k | 423 | 38.1 | 8.5 | 1.8k | 38 | 50.9 | 10.5 | 2k | 42 | 47.2 | 9.6 |
| EN-FR | 15.0k | 264 | 56.9 | 7.7 | 3.0k | 90 | 33.4 | 10.1 | 2k | 65 | 32.2 | 10.1 |

Table 2: Dataset statistics with the number of segments (#seg), number of conversations (#conv), average conversation length (#length), and average number of words per turn (#words) in each split. Note that for KO customer parts, we considered the English reference translation to calculate the number of words.

topics, including account registration issues, payment and delivery clarifications, and after-sale services in various industries such as retail and gaming. The new dataset was automatically anonymized using Unbabel's proprietary anonymization tool, followed by a manual validation performed by expert linguists, to comply with the General Data Protection Regulation (GDPR). The corpus is released under the CC-BY-NC-4.0 license and can be freely used for research purposes only. Please note that, as the license states, no commercial uses are permitted for this corpus.

**Training and Evaluation Datasets.** We provide both training and evaluation (development and test) sets that participants can use to build their systems. Table 2 presents each data splits' statistics, including the number of segments, conversations, and average conversation length. We construct the development and test sets by selecting conversations that exhibit the highest counts of context-dependent discourse phenomena tags, as extracted using Multilingual Discourse Aware (MuDA) tagger (Fernandes et al., 2023b).

### 2.2 Evaluation

We perform a comprehensive evaluation of all submitted systems, using both automatic and human evaluation. Official rankings are determined based on the human assessment scores for both customer and agent translations. We outline the various evaluations conducted below:

#### 2.2.1 Automatic Evaluation

We use COMET (Rei et al., 2022) as our primary evaluation metric for assessing translation quality of the submitted systems.[2] Additionally, we report lexical metrics: BLEU and CHRF using the SacreBLEU library (Post, 2018). We also include CONTEXTCOMETQE (Agrawal et al., 2024), a reference-free metric that uses bilingual context (previous two turns) to assess the translation quality of the current turn. As efficient discourse handling is not directly reflected in standard MT metrics (both lexical and neural), we report the F1 accuracy on the MuDA-tagged discourse phenomena. We considered 4 context-dependent discourse phenomena in our analysis:

- **Lexical cohesion:** Entities may have multiple possible translations in the target language, but the same entity should be referred to by the same word in a conversation.
- **Formality:** Korean uses honorifics to indicate formality, which are special titles or words expressing courtesy or respect for position. In other languages, speakers use second-person pronouns to refer to someone more formally or informally, depending on their relationship with the addressee. Formality should be consistent throughout a conversation.
- **Pronoun resolution:** Some highly inflected languages use gendered pronouns based on semantic or morphological rules. To assign the correct pronoun, it is therefore necessary to use the conversation's context to distinguish the grammatical gender of the pronoun's antecedent.
- **Verb forms:** Verbs must be translated consistently using the form that reflects the tone, and mood of both parties in the conversation.

#### 2.2.2 Manual Evaluation

We use the DA+SQM (Direct Assessment + Scalar Quality Metric) evaluation framework, following the campaigns conducted by the WMT General Translation track over the past years, implemented via the Appraise framework (Federmann, 2018) to collect human assessments of translation quality

[2] Unbabel/wmt22-comet-da

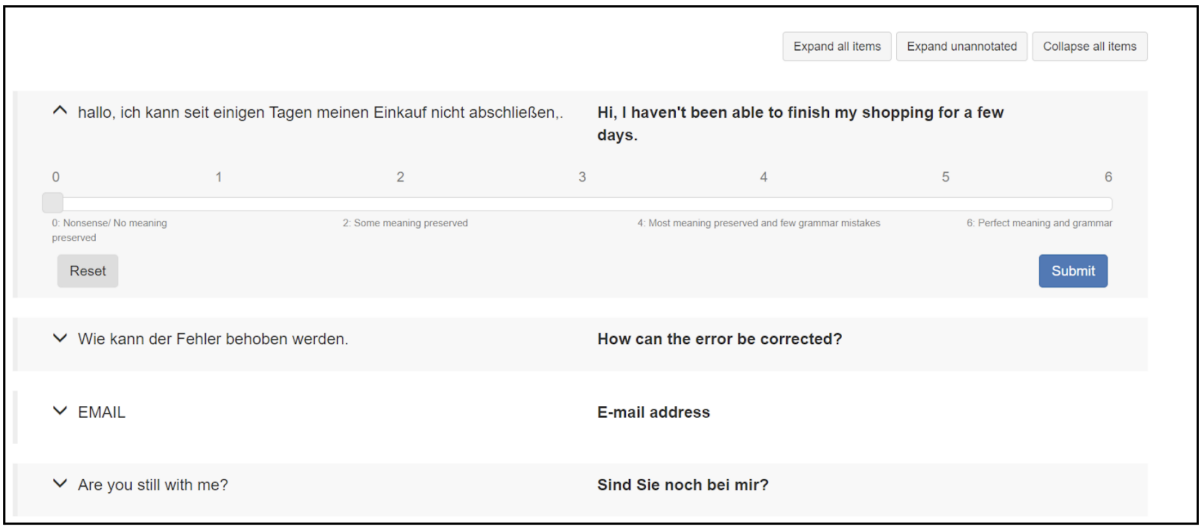


Figure 1: Screen capture of the Appraise interface used by professional linguists to perform human evaluation.

| LP | Threshold | # Chats | # Systems | # annotated segments |
|---|---|---|---|---|
| EN-NL | 35 | 27 | 5 | 3830 |
| EN-PT | 28 | 41 | 5 | 4700 |
| EN-DE | 31 | 36 | 7 | 6629 |
| EN-KO | 48 | 24 | 4 | 3648 |
| EN-FR | 33 | 37 | 6 | 6324 |

Table 3: Statistics of the conversations and instances sampled for the human evaluation step.

for the submitted systems. We ask professional linguists hired via the UpWork[3] platform to evaluate each turn in a conversation within the full context and provide a conversation-level quality score on a continuous scale from 0 to 100. They were instructed to pay special attention to conversation-level properties such as the consistency of style, selection of terms, formality, etc in addition to the correctness criteria. The quality scale includes seven labeled tick marks representing various quality levels based on both accuracy and grammatical correctness (Figure 1).

**Data Selection** For the human evaluation, we retain conversations with up to a given number of turns to make the evaluation manageable. The number of turns for each language pair is specified in Table 3 ("Threshold"), together with the number of conversations and instances retained.

**Measure** We generate turn-level and conversation-level system rankings for each language pair by aggregating the direct assessment scores provided by the linguists at the turn level and the conversation level respectively.

[3] upwork.com

### 2.2.3 LLM-based Error Assessments

LLM-based evaluation has garnered a lot of interest from the community for conducting human-like evaluations. This shift is largely driven by the increasing complexity and scale of language models, making them capable of capturing nuanced understanding and performance of models in real-world tasks. For MT, LLM-based metrics are used to provide fine-grained error assessments over the nature, type, and severity of the errors following the MQM framework (Fernandes et al., 2023a; Lu et al., 2024; Kocmi and Federmann, 2023). Recently, Agrawal et al. (2024) show that context-aware prompting for deriving MQM assessment using LLMs can achieve better correlation with human judgments than the standard MQM prompt for chat translation evaluation, even surpassing COMET.

Hence, we complement our evaluation with an LLM-based fine-grained assessment of MT outputs derived using CONTEXTMQM (Agrawal et al., 2024). The prompt includes the past eight bilingual source sentences as context and one in-domain in-context example with MQM assessment to elicit MQM-like evaluation from GPT-4o-mini[4] for all systems submitted for the EN-DE track.[5] Like MQM, we compute the segment-level *error* score aggregating the number of minor, major, and critical errors, weighted by factors of 10, 5, and 1, respectively.

[4] gpt-4o-mini-2024-07-18 accessed on 10-2-2024.

[5] Due to budget constraints, we conduct this evaluation only on EN-DE, which had the highest number (eight) of participating teams.

## 3 Participants

This section provides a brief description of each participant's systems (§ 3.1). Table 4 summarizes details about the team's institutions and the language directions they participated in. Participants were asked to submit up to three systems per language direction: one primary (explicitly marked) and up to two contrastive systems. Next, we discuss the commonalities and differences between the different submissions § 3.2.

### 3.1 Systems

#### 3.1.1 NLLB-3.3B (Baseline)

For our baseline model, we used the NLLB-3.3B multilingual machine translation model (Costa-jussà et al., 2022) based on an encoder-decoder Transformer architecture (Vaswani et al., 2017). NLLB-3.3B is trained to support over 200 languages, including those of interest in this shared task: English, German, French, Dutch, Brazilian Portuguese, and Korean. We opted for a sentence-level baseline that does not incorporate additional context and used a beam size of 4 for generating translation hypotheses.

#### 3.1.2 UNBABEL-IT

The joint submission of Unbabel and IT includes one primary submission and two contrastive submissions per language pair. The systems are based on Tower-7B models and are trained on the chat datasets released by the shared task. Their primary system uses contextual MBR re-ranking over a set of 50 candidates to get the best hypothesis. Additionally, the first contrastive submission is a 70B variant of the Tower model specialized to have general purpose translation capabilities and the second one uses greedy decoding with the 7B model finetuned on chat datasets.

#### 3.1.3 DEEPTEXT LAB

DEEPTEXT LAB participated in the English-Korean language pair with a single primary system. Their submission leverages Google's Gemma-2-27B model [6], using the most recent two turns and summaries of previous turns as context, all within the same document. The turn summaries are generated using the GPT-4o-mini model. Their system was trained solely using the training data provided by the shared task.

| TEAM | INSTITUTION | DIRECTIONS |
|---|---|---|
| DeepText Lab | Yonsei University | EN-KO |
| HW-TSC | Huawei Translation Service Center | EN-DE |
| Multitan-GML | Université Paris Cité | EN-FR |
| SETU-ADAPT | ADAPT research centre & Dublin City University | EN-DE, EN-FR |
| SheffieldGATE | University of Sheffield | EN-DE, EN-NL, EN-PT |
| CLTeam | Vrije Universiteit Amsterdam | EN-DE, EN-NL, EN-FR, EN-PT |
| DCUGenNLP | Dublin City University | ALL |
| Unbabel-IT | Unbabel & Instituto de Telecomunicações | ALL |
| Baseline | Organizers | ALL |

Table 4: The participating teams, their affiliations, and the language directions that they participated.

#### 3.1.4 HW-TSC

Huawei Translation Service Center (HW-TSC) team submitted a primary and two contrastive systems for English↔German language pair. Their system is a 25-6 transformer encoder-decoder model with a feed-forward dimension of 4096 and 16 self-attention layers. Their primary submission uses a model from the previous edition of the shared task as a baseline, finetuned on this edition's training data, followed by a second finetuning on the validation data. Next, they use MBR reranking to select the optimal candidate with COMET as the utility function using outputs generated from a diverse set of models. Their system then undergoes a self-training step on the MBR output. The contrastive submissions include models trained with different finetuning strategies (e.g. excluding the finetuning on the dev set).

#### 3.1.5 SHEFFIELDGATE

The SHEFFIELDGATE team participated in English↔German, English↔Dutch, and English↔Brazilian Portuguese, with one primary system per language pair. Their system performs low-rank (Hu et al., 2022) instruction-tuning with the training and validation datasets provided by the shared task on the Llama-3-8B-Instruct [7] model. To incorporate contextual information and dependencies between chat messages, they introduce a context-aware sliding window approach that incorporates translations generated at each turn into the prompt.

[6] google/gemma-2-27b-it

[7] meta-llama/Meta-Llama-3-8B-Instruct

| PARTICIPANT | BASE MODEL | CHAT CONTEXT? | IN-DOMAIN TRAINING? | MULTILINGUAL? | SYNTHETIC DATA? | DECODING |
|---|---|---|---|---|---|---|
| DeepText Lab | Gemma-2-27B | ✓(summary) | ✓ | ✗ | ✗ | NR |
| HW-TSC | Transformer 25-6 (from scratch) | ✗ | ✓ | ✗ | ✓ | MBR |
| Multitan-GML | Commercial* | ✓ | ✓ | ✗ | ✗ | NR |
| SETU-ADAPT | Llama-3-8B (EN-DE) | ✓(few-shot) | ✓ | ✗ | ✗ | NR |
| | NLLB-200-600M (EN-FR) | ✗ | ✓ | ✗ | ✓ | NR |
| SheffieldGATE | Llama-8b-Instruct | ✓ | ✓ | ✓ | ✗ | NR |
| CLTeam | TowerInstruct-7B-v0.2 | ✓(graph) | ✗ | ✓ | ✗ | NR |
| DCUGenNLP | Llama3.1-8b | NR | ✓ | ✓ | ✗ | NR |
| Unbabel-IT | TowerBase-7B | ✓ | ✓ | ✓ | ✗ | MBR |
| Baseline | NLLB-3.3B | ✗ | ✗ | ✓ | ✗ | Beam (4) |

Table 5: Summary of approaches for all primary submissions. NR: Not reported.

#### 3.1.6 SETU-ADAPT

SETU-ADAPT team submitted 3 (one primary and two contrastive) systems based on different pre-trained models: NLLB[8], MBART-50[9] and Llama-3-8B[10]. Their primary system for EN-DE uses a Llama-3-8B backbone finetuned on the in-domain chat and a synthetic dataset generated by back-translating domain-specific monolingual sentences. For EN-FR, they finetune an NLLB-600M model. During inference, with the LLM-based models, they perform few-shot prompting using examples retrieved via similarity search from the training dataset. Their contrastive systems are based on the encoder-decoder models but use the same datasets for training.

#### 3.1.7 MULTITAN-GML

MULTITAN-GML's primary system finetunes a "Dialog" in-domain specialized model hosted on the Model Studio Lite server [11] with 2022 Chat Task (train, valid, test) and 2024 Chat Task (valid) datasets. Their two contrastive submissions use outputs from NLLB-3.3B model and the Deep_translator API respectively. All outputs are post-edited using GPT-4o.

#### 3.1.8 DCUGENNLP

DCUGENNLP team submitted a total of 15 systems (one primary and two contrastive) for all the five language pairs. Their primary system finetunes a Llama-3.1-8B model on a mix of the chat task's training data and datasets from other WMT tracks. They also include synthetically generated customer-service data generated using one of their contrastive submission. Other contrastive submissions use Mistral-7B as base models with optional prompt tuning or finetuning of adapter layers.

#### 3.1.9 CLTEAM

CLTEAM submitted one primary and one contrastive systems for each of the English↔German, English↔French, English↔Dutch, and English↔Brazilian Portuguese language pairs. Their system uses TowerInstruct-7B-v0.2 [12] model as the base LLM. For their primary submission, they prompt the model with both the dialogue history represented using a graph and the source sequence to be translated. To generate the graph, they prompt GPT-4o to extract entities and relationships from the dialogue data, creating triples from these elements. For the contrastive submission, they prompt the model with only the source sequence to be translated.

### 3.2 Discussion

Table 5 presents a summary of approaches used by all the submitted systems. We highlight some key aspects below:

**Model Architecture** Most teams except CLTEAM and HW-TSC finetuned general-purpose pre-trained LLMs. Where CLTEAM used an off-the-shelf translation-finetuned LLM, HW-TSC opted for a custom bilingual encoder-decoder model for their participation.

**Training Data** All teams used the provided training and development data, sourced from the current and previous versions of the task. HW-TSC went a step further by generating a synthetic parallel corpus. They did this by forward translating source-side monolingual data into target-side text and backtranslating target-side monolingual into source-side texts. SETU-ADAPT similarly used

[8] facebook/nllb-200-distilled-600M
[9] facebook/mbart-large-50-many-to-many-mmt
[10] unsloth/llama-3-8b-bnb-4bit
[11] modelstudio-lite
[12] Unbabel/TowerInstruct-7B-v0.2

| System | EN-DE | | EN-FR | | EN-NL | | EN-PT | | EN-KO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | FR | EN | NL | EN | PT | EN | KO | EN |
| DeepText Lab | | | | | | | | | 93.03 | 94.11 |
| HW-TSC | **93.58** | **93.30** | | | | | | | | |
| MULTITAN-GML | | | 90.09 | 92.42 | | | | | | |
| ADAPT | 90.59 | 90.97 | 82.19 | 82.69 | | | | | | |
| SheffieldGATE | 88.67 | 90.10 | | | 88.93 | 89.71 | 90.05 | 88.12 | | |
| CLTeam | 90.90 | 91.63 | 91.37 | 91.90 | 91.31 | 91.22 | 91.77 | 90.12 | | |
| DCUGenNLP | 90.49 | 91.10 | 91.05 | 90.73 | 91.32 | 90.96 | 93.24 | 89.66 | 91.50 | 93.41 |
| Unbabel-IT | 93.22 | 92.48 | **92.96** | **92.71** | **94.36** | **93.38** | **94.76** | **92.46** | **94.96** | **95.16** |
| NLLB-3.3B | 90.56 | 89.03 | 91.06 | 89.18 | 87.86 | 88.45 | 86.33 | 86.10 | 87.26 | 88.05 |
| Δ (Best) | +3.02 | +4.27 | +1.9 | + 3.53 | +6.50 | +4.93 | +8.43 | +6.36 | +7.70 | +7.11 |

Table 6: COMET results on the official test set. Δ (Best): improvement over baseline.

| System | EN-DE | | EN-FR | | EN-NL | | EN-PT | | EN-KO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | FR | EN | NL | EN | PT | EN | KO | EN |
| DeepText Lab | | | | | | | | | 57.67 | 77.96 |
| HW-TSC | **82.66** | **84.03** | | | | | | | | |
| MULTITAN-GML | | | 79.54 | **82.71** | | | | | | |
| ADAPT | 69.50 | 76.63 | 63.92 | 55.98 | | | | | | |
| SheffieldGATE | 64.94 | 72.04 | | | 60.01 | 68.31 | 67.67 | 66.38 | | |
| CLTeam | 69.87 | 75.39 | 74.66 | 77.41 | 63.59 | 73.00 | 71.38 | 69.45 | | |
| DCUGenNLP | 69.84 | 73.64 | 73.73 | 73.78 | 67.44 | 70.47 | 75.24 | 67.27 | 49.02 | 75.35 |
| Unbabel-IT | 77.23 | 79.87 | **80.51** | 78.57 | **80.25** | **78.60** | **82.55** | **76.01** | **62.29** | **81.57** |
| NLLB-3.3B | 70.22 | 71.79 | 76.03 | 76.37 | 59.55 | 68.62 | 58.60 | 67.13 | 34.50 | 69.87 |
| Δ (Best) | +12.44 | +12.24 | +4.48 | +6.34 | +20.70 | +9.98 | +23.95 | +8.88 | +27.79 | +11.70 |

Table 7: CHRF results on the official test set. Δ (Best): improvement over baseline.

back translation to generate more in-domain data for their EN-FR submission.

**Inference** Both UNBABEL-IT and HW-TSC leveraged a quality-aware decoding (QAD) approach (Fernandes et al., 2022) for further improving the quality of outputs during inference. While HW-TSC optimized for COMET, UNBABEL-IT used a context-aware COMET metric as a utility for selecting the best candidate. HW-TSC also used MBR outputs to further finetune the model.

**Context Usage** Different strategies were employed to incorporate conversation context into the translation process. UNBABEL-IT, SHEFFIELDGATE, and MULTITAN-GML utilized the previous turns of the conversation as context to maintain continuity and coherence in translations. DEEPTEXT LAB used both the previous two turns as well as the summary of all the previous conversation turns except the last two, generated by GPT-4o-mini. This allowed the model to focus on the essential part of the previous content without being overwhelmed by excessive details. On the other hand, CLTEAM used a graph representation of the conversation's history as context, capturing the connectivity between various concepts thus serving as a compressed memory of the dialogue context. SETU-ADAPT used few shot examples extracted from the training data using sentence-embedding similarity.

All teams that participated for more than one language pair opted for a multilingual system except for SETU-ADAPT team who submitted two different systems for each language pair they participated in (EN-DE, EN-FR).

## 4 Overall Results

We present the results of the automatic evaluation for all participating systems for all language pairs

| System | EN-DE | | EN-FR | | EN-NL | | EN-PT | | EN-KO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | FR | EN | NL | EN | PT | EN | KO | EN |
| DeepText Lab | | | | | | | | | 15.99 | 16.15 |
| HW-TSC | 20.79 | 23.37 | | | | | | | | |
| MULTITAN-GML | | | 0.31 | 0.21 | | | | | | |
| ADAPT | 15.63 | 17.97 | -23.31 | -22.88 | | | | | | |
| SheffieldGATE | 17.87 | 17.54 | | | 13.72 | 14.39 | 5.87 | 3.46 | | |
| CLTeam | 19.65 | 21.15 | 8.22 | 7.26 | 19.00 | 19.20 | 8.64 | 7.68 | | |
| DCUGenNLP | 17.27 | 20.38 | 5.11 | 4.80 | 16.55 | 16.09 | 8.69 | 6.70 | 15.84 | 15.74 |
| Unbabel-IT | **24.41** | **26.15** | **10.67** | **10.00** | **23.93** | **23.39** | **12.74** | **10.59** | **21.64** | **21.08** |
| NLLB-3.3B | 15.56 | 19.09 | 1.24 | 0.77 | 9.35 | 8.04 | -5.51 | -6.75 | 4.11 | 4.13 |
| Δ (Best) | +8.85 | +7.06 | +9.43 | +9.23 | +14.58 | +15.35 | +18.25 | +17.34 | +17.53 | +16.95 |

Table 8: ContextCometQE results on the official test set. Δ (Best): improvement over baseline.

| System | EN-DE | | EN-FR | | EN-NL | | EN-PT | | EN-KO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | FR | EN | NL | EN | PT | EN | KO | EN |
| DeepText Lab | | | | | | | | | 37.65 | 66.98 |
| HW-TSC | **68.76** | **71.27** | | | | | | | | |
| MULTITAN-GML | | | 65.43 | **71.80** | | | | | | |
| ADAPT | 51.39 | 59.90 | 33.17 | 28.56 | | | | | | |
| SheffieldGATE | 41.15 | 50.72 | | | 33.62 | 46.54 | 42.58 | 42.25 | | |
| CLTeam | 50.41 | 55.71 | 57.05 | 61.09 | 39.29 | 55.41 | 46.42 | 49.34 | | |
| DCUGenNLP | 49.97 | 57.29 | 56.32 | 56.39 | 46.38 | 52.15 | 56.36 | 45.87 | 27.66 | 62.13 |
| Unbabel-IT | 61.45 | 62.86 | **66.41** | 63.18 | **65.70** | **63.75** | **67.86** | **59.05** | **41.54** | **71.01** |
| NLLB-3.3B | 50.43 | 52.09 | 59.21 | 58.07 | 33.55 | 48.47 | 28.25 | 45.59 | 12.46 | 49.76 |
| Δ (Best) | +18.33 | +19.18 | +7.20 | +13.73 | +32.15 | +15.28 | +39.61 | +13.46 | +29.08 | +21.25 |

Table 9: BLEU results on the official test set. Δ (Best): improvement over baseline.

in § 4.1. We then discuss findings from human evaluation in § 4.2, followed by an LLM-based error assessment of submitted systems for the EN-DE task in § 4.3.

### 4.1 Automatic Evaluation

Tables 6-9 show the results of automatic evaluations on the official test set using Comet, chrF ContextCometQE and BLEU respectively -- most participant systems improve the translation quality according to both neural (Comet, ContextCometQE) and lexical (chrF, BLEU) metrics over the NLLB-3.3B model, except the SETU-ADAPT system for EN-FR. This can be explained by the fact that SETU-ADAPT finetunes an NLLB-600M model for EN-FR, which, albeit from the same family of models as our baseline (NLLB-3.3B), is significantly smaller in size.

The Unbabel-IT submission consistently outperforms all other systems, except the EN-DE translation task, where the winning submission according to Comet, BLEU, and chrF is HW-TSC. Similarly, Multitan-GML scores the best on BLEU and chrF when translating French into English. Interestingly both systems (Unbabel-IT and HW-TSC) use MBR decoding with ContextComet and Comet respectively, suggesting that inference optimization techniques like quality-aware decoding methods (Fernandes et al., 2022) can be useful in pushing the translation quality of strong MT systems. However, as we will see in §4.2, this difference is not reflected in human assessments and in automatic metrics (ContextMQM and ContextCometQE), with different methods scoring the two systems differently. This highlights the importance of carefully selecting the optimized metrics and the evaluation criteria, as over-optimizing certain metrics may lead to mixed or misleading outcomes (Fernandes et al., 2022).

Unbabel-IT's submission also achieves the highest scores across all settings according to ContextCometQE. However, we observe that the range of quality scores produced by the ContextCometQE model, when aggregated at the system level, significantly deviates from the typical range of this metric.[13] While Agrawal et al. (2024) demonstrate its effectiveness as a segment-level

[13] System-level scores are higher when the context is not considered.

metric with improved correlation to human judgments, further investigation is necessary to understand how these system-level scores should be interpreted. For instance, MULTITAN-GML, which performs well on lexical metrics such as BLEU and CHRF, receives a notably lower score with CONTEXTCOMETQE.

| System | Precision | Recall | F1 |
|---|---|---|---|
| HW-TSC | 76.7 | 86.2 | 81.2 |
| SETU-ADAPT | 75.0 | 69.2 | 72.0 |
| SheffieldGATE | 73.0 | 70.8 | 71.9 |
| CLTeam | 75.7 | 81.5 | 78.5 |
| DCUGenNLP | 74.6 | 81.5 | 77.9 |
| Unbabel-IT | 75.4 | 66.2 | 70.5 |
| NLLB-3.3B | 74.3 | 84.6 | 79.1 |

Table 10: MUDA scores for EN-DE pronouns.

**Discourse Phenomena Analysis** Figure 2 shows the F1 accuracy for all systems in correctly using the discourse markers across multiple phenomena for all language pairs. The baseline system (NLLB-3.3B) has competitive accuracy with submitted systems on higher resource language pairs (EN→DE and EN→FR). For all settings except "pronouns" for German and "formality" for German and French, UNBABEL-IT achieves the highest accuracy across the board. Surprisingly, the MUDA F1 score for correctly generating German pronouns is worse for UNBABEL-IT relative to the baseline. A qualitative analysis shows that this is due to pronouns being under-generated in UNBABEL-IT's translations resulting in high precision but low recall scores as shown in Table 10.

To validate the observations and findings derived from automatic metrics, we now turn to human evaluation of the submitted systems for a more reliable assessment of translation quality.

## 4.2 Human Evaluation

We present the human evaluation results at both turn and conversation levels in Tables 11 and 12 respectively.

**Overall results.** UNBABEL-IT outperforms all systems on both turn-level and conversation-level evaluation, surpassing the HW-TSC system that achieved the highest COMET scores on EN-DE

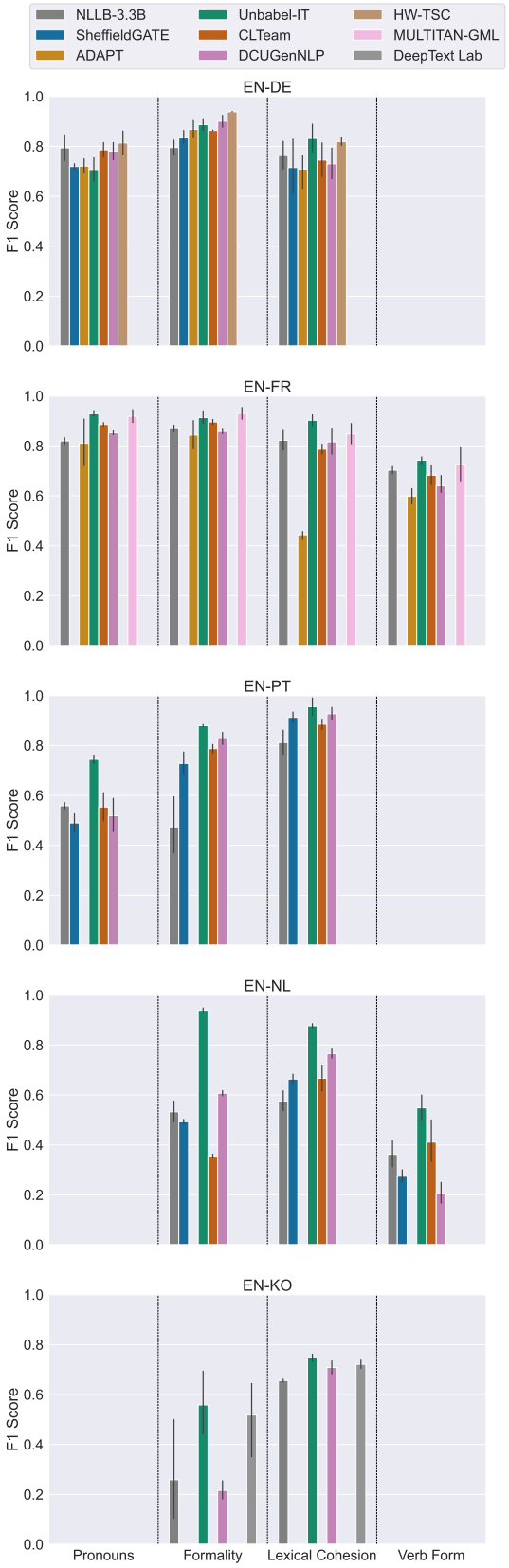


Figure 2: MUDA F1 scores across all settings.

translation pair.[14] The translation quality according

[14] We note that the human evaluation for EN-DE, like other LPs, was conducted on a subset of the dataset (limited to a maximum of 30 turns per conversation).

| System | EN-DE | | EN-FR | | EN-NL | | EN-PT | | EN-KO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | XX | EN | XX | EN | XX | EN | XX | EN | XX | EN |
| DeepText Lab | | | | | | | | | 91.35 | 95.71 |
| HW-TSC | 88.47 | 90.41 | | | | | | | | |
| MULTITAN-GML | | | 81.83 | 84.62 | | | | | | |
| ADAPT | 82.55 | 88.83 | 70.22 | 65.53 | | | | | | |
| SheffieldGATE | 78.63 | 88.85 | | | 85.62 | 94.18 | 73.34 | 81.53 | | |
| CLTeam | 83.12 | 89.12 | 84.28 | 85.79 | 93.39 | 95.83 | 74.14 | 80.52 | | |
| DCUGenNLP | 84.56 | 88.60 | 85.72 | 83.26 | 91.30 | 94.61 | 80.21 | 81.55 | 89.71 | 96.15 |
| Unbabel-IT | **89.42** | **92.74** | **90.24** | **90.00** | **98.16** | **97.40** | **82.04** | **82.37** | **93.39** | **96.31** |
| NLLB-3.3B | 78.05 | 87.57 | 80.59 | 77.82 | 82.66 | 90.98 | 61.27 | 73.98 | 79.13 | 90.47 |

Table 11: Human Evaluation results aggregated at the turn level on the official test set.

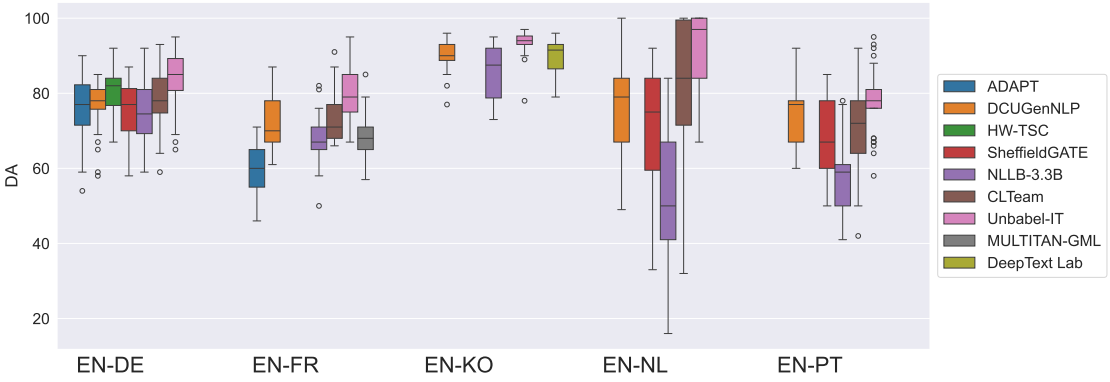


Figure 3: Conversation-level DA scores.

| | EN-DE | EN-FR | EN-NL | EN-PT | EN-KO |
|---|---|---|---|---|---|
| DeepText Lab | | | | | 90.04 |
| HW-TSC | 81.19 | | | | |
| MULTITAN-GML | | 68.59 | | | |
| ADAPT | 75.75 | 59.65 | | | |
| SheffieldGATE | 75.72 | | 70.81 | 68.27 | |
| CLTeam | 78.61 | 73.32 | 84.37 | 69.85 | |
| DCUGenNLP | 77.03 | 72.27 | 76.41 | 73.78 | 89.83 |
| Unbabel-IT | **84.22** | **79.62** | **92.22** | **78.00** | **93.21** |
| NLLB-3.3B | 74.50 | 67.81 | 53.07 | 56.37 | 85.63 |

Table 12: Human Evaluation results aggregated at the conversation level on the official test set.

to direct assessment scores of all systems evaluated across all language pairs is high ($> 65$) at both conversation and turn levels. This could be because of the nature of the chat dataset which contains very short texts (the number of words per turn across language pairs is less than 8, see Table 2).

**Conversation-level results.** Figure 3 shows the distribution of scores assigned at the conversation level for all systems and language pairs. Confirming the automatic results, NLLB-3.3B scores the lowest and with the highest standard deviation for EN-KO, EN-NL and EN-PT. We also observe that EN-NL generally exhibits the largest standard deviation. After analyzing the outputs, we found that EN-NL has the highest number of segments (and conversations) receiving either a score of 0 (when hallucinating or copying source text verbatim) or 100, indicating a significant variation in translation quality for this language pair. Although there are sentences with mid-range scores, the dominance of segments with extremely high or low scores greatly influences the overall results, substantially raising the standard deviation.

**Turn-level results.** Figure 4 illustrates DA scores with the increase in the number of turns. For most systems and language pairs, translation quality deteriorates over successive turns, indicating a decline in the systems' ability to maintain consistency and accuracy in prolonged dialogues. This decline is particularly evident in the baseline sys-

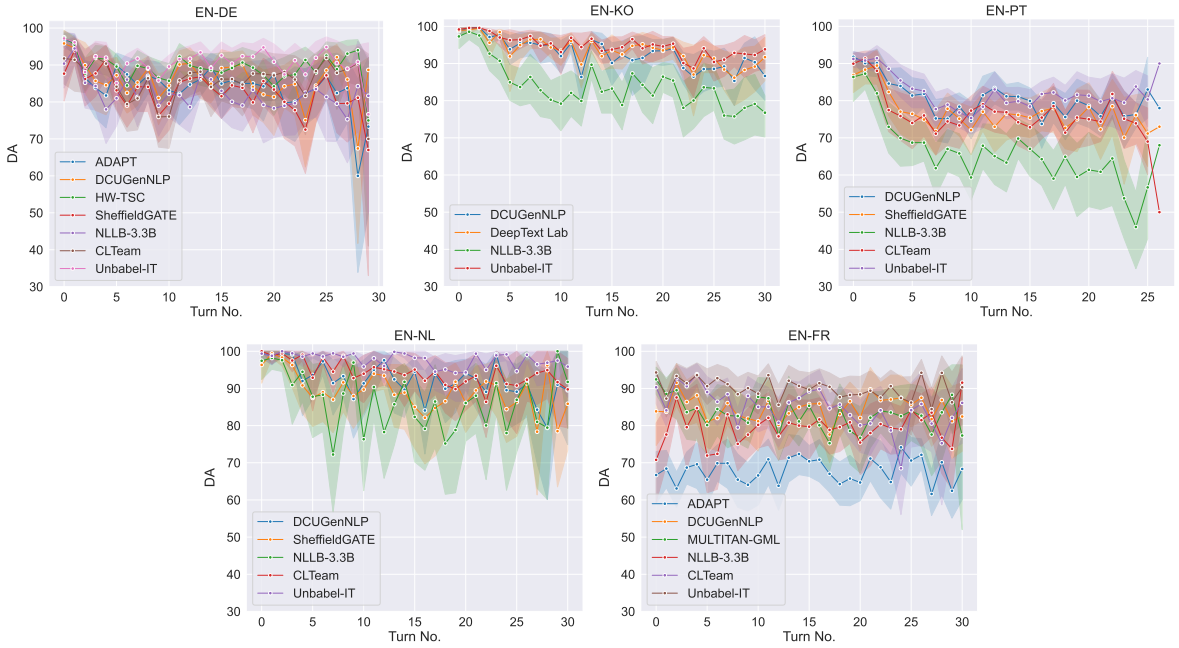


Figure 4: Turn-level DA score across different language pairs through a chat.

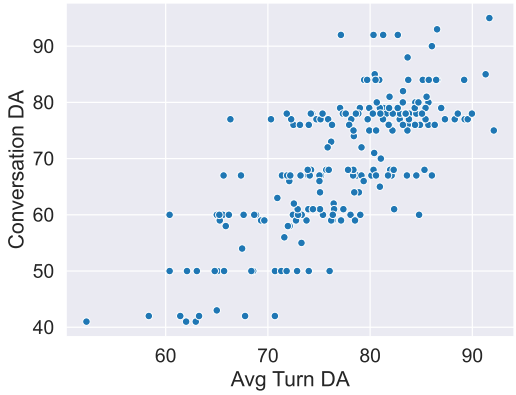


Figure 5: Turn avg. vs conversation-level DA scores.

tem, which does not leverage contextual information from previous turns to generate translations. Interestingly, however, despite not using contextual information, HW-TSC's system maintains translation quality across successive turns. This can likely be attributed to rigorous training on in-domain data, both authentic and synthetically generated.

**Turn Vs. Conversation Quality results.** Overall, **conversation-level quality is lower than turn-level scores** suggesting that there are aspects beyond translation accuracy that might impact the overall translation quality and user experience. This is corroborated by the observation that the Spearman correlation between the average turn-level score and conversation-level DA score, though high, is 0.722. For future evaluations, it might be worth investigating dialogue-oriented human assessment (Mendonca et al., 2023) to understand how turn-level scores impact conversation-level quality.

While direct assessments from experts provide a reliable measure of translation quality, DA scores fall short in offering insights into when and how errors occur, as well as their types and nature. Therefore, to assess the severity of errors generated by these systems, we now turn to LLM-based fine-grained error assessment of translation outputs.

### 4.3 LLM-based Evaluation

| System | % Perfect | # Minor | # Major | # Critical | Avg. Score |
|---|---|---|---|---|---|
| HW-TSC | 89.12 | 100 | 88 | 59 | -0.554 |
| SETU-ADAPT | 82.61 | 158 | 139 | 99 | -0.903 |
| SheffieldGATE | 77.95 | 220 | 178 | 95 | -1.009 |
| CLTeam | 86.28 | 139 | 82 | 79 | -0.656 |
| DCUGenNLP | 83.10 | 143 | 158 | 80 | -0.849 |
| Unbabel-IT | **94.41** | **51** | **47** | **18** | **-0.228** |
| NLLB-3.3B | 80.50 | 161 | 143 | 117 | -1.002 |

Table 13: CONTEXTMQM scores for EN-DE.

Table 13 shows the results from using LLM-based error assessments via CONTEXTMQM. UNBABEL-IT leads the pack with 94.41% perfect translations. It also has the lowest number of errors in each category (minor, major, and critical), with an average error score of -0.228 (less than 1 minor error), the best among all systems. All systems, however, manage to achieve over 77% perfect

translations, meaning the overall quality across the board is strong.

Despite the positive results, there are notable differences in error distribution. For example, both the SHEFFIELDGATE and SETU-ADAPT models, while maintaining a reasonable percentage of perfect translations (82.61% and 77.95%, respectively), suffer from a significantly higher number of errors across all categories—minor, major, and critical. This suggests that when these systems do make errors, they tend to be more frequent and more serious, dragging down their overall performance compared to other systems. Interestingly, contrary to human evaluation but in line with other automatic measures, DCUGENNLP scores worse than CLTEAM submission, highlighting limitations of existing evaluation methods to discern systems with close translation quality.

## 5 Conclusions

This paper presents the findings of the Chat Translation Shared Task 2024. This year, we expanded the set of language pairs to include two additional languages (EN-KO and EN-NL). We created the evaluation sets with a focus on context usage when assessing system performance. We also employed a range of complementary evaluation methods to assess all systems, including automatic metrics that focus on translation quality, as well as fine-grained error assessments and analysis of specific discourse phenomena.

We find that the best systems finetune strong pretrained LLMs using multilingual in-domain data and use contextual information (such as graphs, summaries or raw context) during training and inference. Additionally, using synthetic data during training improved translation quality. Furthermore, QAD strategies were effective in aligning translations with quality expectations.

As future work, a possible direction is to leverage reference-free discourse quality metrics that can give complementary insights to the translation evaluation approaches we tried this year. It might also be worth investigating human and automatic evaluation frameworks that assess specific dimensions relevant to chat (e.g. fluidity, coherence, consistency, etc).

## Limitations

Due to budget constraints, we conducted human evaluations using DA on a subset of the test set, which limited the number of turns evaluated for each language pair. For similar cost-related reasons, we ran CONTEXTMQM on a single language pair that received the highest number of submissions. Additionally, we note that our analysis of discourse-specific phenomena is constrained by the quality of taggers, which only annotate specific properties based on predefined rules and may not fully capture all levels of ambiguity present in chat datasets.

## Ethics Statement

**Data released.** MAIA 2.0 corpus includes conversations from clients who gave written consent to use this data for research purposes as long as it follows the General Data Protection Regulation (GDPR). The original segments of customers and agents are translated via the MTPE (Machine Translation followed by a Post-Editing) process by experienced translators of the Unbabel Community. To make the data publicly available, the corpus was first anonymized automatically by using the Unbabel proprietary anonymization tool and then went through manual verification.

**Human evaluation.** Human direct assessment of system outputs was performed by professional translators hired via the UpWork platform and paid at professional rates.

**Reproducibility.** For LLM-based error assessment, the continual updates of closed commercial models present challenges to the reproducibility of research. We have included the exact version of the model we used, along with the precise date of evaluation.

## Acknowledgements

This work was supported by the 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

Sujan Reddy A. 2022. Automating human evaluation of dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*,

pages 229–234, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Sweta Agrawal, M. Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Is context helpful for chat translation evaluation? *CoRR*, abs/2403.08314.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

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.

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.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

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

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. 2023b. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

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

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8801–8816, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.