Unlocking Latent Discourse Translation in LLMs Through Quality-Aware Decoding

Wafaa Mohammed[1] **Vlad Niculae**[1] **Chrysoula Zerva**[2,3]
[1]University of Amsterdam
[2]Instituto Superior Técnico, University of Lisbon
[3]Instituto de Telecomunicações
{w.m.a.mohammed, v.niculae}@uva.nl, chrysoula.zerva@tecnico.ulisboa.pt

Abstract

Large language models (LLMs) have emerged as strong contenders in machine translation. Yet, they still struggle to adequately handle discourse phenomena, such as pronoun resolution and lexical cohesion at the document level. In this study, we thoroughly investigate the discourse phenomena performance of LLMs in context-aware translation. We demonstrate that discourse knowledge is encoded within LLMs and propose the use of quality-aware decoding (QAD), specifically minimum Bayes risk decoding, to effectively extract this knowledge, showcasing its superiority over other decoding approaches through comprehensive analysis. Furthermore, we illustrate that QAD enhances the semantic richness of translations and aligns them more closely with human preferences.

1 Introduction

Large language models (LLMs) have demonstrated superior performance in machine translation (MT), producing strong results for sentence-level and document-level translation (Wang et al., 2023; Xu et al., 2023; Alves et al., 2024; Zhu et al., 2024). Quality improvements in document-level translation are key in producing translations that align better with human preferences, since documents are the natural way in which we consume and produce text (Läubli et al., 2018; Maruf et al., 2022; Mohammed and Niculae, 2024b; Dahan et al., 2024). Additionally, document-level translation provides a means to tackle discourse-related challenges in translation, including inter-sentential coreference resolution as well as the need for maintaining coherence, style, and formality level across the document (Post and Junczys-Dowmunt, 2023).

At the same time, it has been observed that LLM-derived translations frequently feature different linguistic and semantic characteristics and patterns compared to neural machine translation (NMT), hence inspiring several works that try to trace and understand such patterns and differences. Thus, recent work ranges from designing linguistic performance test suites (Manakhimova et al., 2024) to analyzing specific aspects such as lexical features, literalness, formality (Wisniewski et al., 2024), gender bias (Kotek et al., 2023; Zhao et al., 2024), and pronoun resolution. These studies uncovered valuable features of LLM-derived translations, including suboptimal performance compared to NMT systems in several phenomena, such as punctuation, future verb tenses, stripping, function words (Manakhimova et al., 2024), and pronoun resolution (Mohammed and Niculae, 2024a). Other works observed that LLMs show systematic differences to NMT systems in their choice of lexical features, such as part-of-speech (PoS) patterns (Sizov et al., 2024) as well as their ability to produce less literal translations while remaining competitive quality-wise to NMT translations (Raunak et al., 2023).

Despite these insights, fine-grained analyses rarely extend to document-level MT, where discourse context makes such phenomena even more critical and further underscores the need to understand the linguistic and semantic properties of LLM translations. We thus aim to study LLMs' performance in document-level MT (particularly, context-aware MT) with respect to different discourse phenomena. Inspired by Fernandes et al. (2023), we measure models' performance on four phenomena: lexical repetition, pronoun resolution, formality, and verb forms. We compare the performance of recent translation-LLMs to encoder-decoder models on the DELA corpus, a high-quality human-curated dataset that is rich in discourse phenomena (Castilho et al., 2021). Moreover, we hypothesize that discourse knowledge can be implicitly encoded in LLMs, but is not fully exploited by greedy decoding. Following prior work (Shin et al., 2020; Kojima et al., 2022), we adopt an inference-time probing approach to test this hypothesis. We thus experiment with quality-aware decoding (Fer-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4752–4774
March 24-29, 2026 ©2026 Association for Computational Linguistics

Lexical repetition	**EN:** The reviewer gave us constructive feedback. We appreciate the reviewer 's feedback. **FR:** L'examinatrice nous a fait un retour constructif. Nous apprécions le retour de l' examinatrice .
Pronoun resolution	**EN:** One of the Chinese worked in an amusement park. It was closed for the season. **DE:** Ein Chinese arbeitete in einem Vergnügungspark. Er war gerade geschlossen.
Formality	**EN:** How are you my dear friend? Would you like to go to the cinema with me? **DE:** Wie geht es dir, mein lieber Freund? Möchtest du mit mir ins Kino gehen?
Verb form	**EN:** Maria said she was too sick. However, she was seen walking in the park. **PT:** A Maria disse que estava muito doente. No entanto, ela foi vista a passear no parque.

Table 1: Examples of discourse phenomena. Ambiguous words are highlighted in pink , and supporting context necessary to resolve the ambiguity is marked in underlined purple text.

nandes et al., 2022), specifically minimum Bayes risk (MBR) decoding (Eikema and Aziz, 2020), and find that it indeed helps improve the discourse phenomena performance of LLMs. We validate our findings through extensive experiments on six language pairs from three language families: English to Brazilian-Portuguese, German, French, Korean, Arabic and Russian, on two datasets, namely, TED2020 (Reimers and Gurevych, 2020) and WMT24++ dataset (Deutsch et al., 2025). Moreover, we perform an ablation study on different inference setups, including MBR decoding with different choices for the utility function incorporating discourse specific metrics (Wong and Kit, 2012; Zhao et al., 2023), automatic post editing, and sample fusion (Vernikos and Popescu-Belis, 2024). We note that our approach provides empirical evidence for leveraging discourse-relevant information in LLMs; however, internal mechanistic analyses are required to fully substantiate the hypothesis that discourse knowledge is encoded in these models. We hope our work serves as an initial step towards this goal.

Our contributions can be summarized as follows:

- We design a comprehensive evaluation setup leveraging a discourse-rich dataset, showing that under greedy decoding, encoder-decoder models outperform LLMs in discourse performance.
- We demonstrate through extensive evaluation on six language pairs using automatic metrics, LLM-as-a-judge, and human assessment that QAD improves the translation and the discourse performance of LLMs, enabling them to surpass encoder-decoders.
- We conduct a comprehensive analysis on the effect of different inference setups on discourse performance.
- We release human annotations based on TED2020 that focus on discourse phenomena, supporting further research in this area.[1]

2 Background

2.1 Discourse Phenomena in Document-Level Translation

Translating beyond the sentence level brings extra challenges that concern inter-sentential coreference resolution, lexical repetition, and coherence. Handling these challenges is important to ensure reliable, adequate translations that align with human preferences. In this work, we focus on four linguistic phenomena that are relevant to document-level translation, in which the required contextual information extends beyond sentence-level boundaries, as proposed by (Fernandes et al., 2023):

Lexical repetition. Entities mentioned multiple times in a document should be translated in the same way.

Pronoun resolution. For languages that have gendered pronouns, the translation should respect the gender of the referent.

Formality. In some languages, linguistic indicators such as pronouns and honorifics are used when addressing someone formally or expressing respect. The same level of formality should be maintained across a document.

Verb form. In some languages, verb forms vary depending on the subject, tense, and consistency with the context. For instance, in Arabic, the verb form changes according to whether the subject is singular, dual, or plural, masculine or feminine. Verb forms should be accurate across a document.

Examples of the phenomena are presented in Table 1. The discourse phenomena we study here are for EN→XX translation direction. Studying the

[1] Code and data are at https://github.com/Wafaa014/Discourse_translation_in_LLMs/.

reverse direction (XX →EN) requires a separate set of analyses to determine the relevant phenomena in that case, which is out of scope of this work.

2.2 Quality-Aware Decoding (QAD)

Quality-aware decoding for machine translation refers to utilizing translation evaluation metrics during decoding to choose the best candidate among several sampled responses from the model. Samples can be generated using vanilla temperature sampling or variations of it that truncate the distribution, such as top-k or nucleus sampling (Fan et al., 2018; Holtzman et al., 2020). QAD has been proven to generate better quality translations compared to maximum-a-posteriori (MAP) decoding according to automatic metrics and human evaluation (Fernandes et al., 2022). There are different approaches to quality aware-decoding including reranking (Lee et al., 2021; Bhattacharyya et al., 2021), minimum Bayes risk (MBR) decoding (Eikema and Aziz, 2020, 2022; Müller and Sennrich, 2021), and fusion of samples (Vernikos and Popescu-Belis, 2024). In our main experiments, we focus on MBR since it is a widely used approach for MT with proven benefits on overall translation performance (Wu et al., 2024b; Li et al., 2024a; Kudo et al., 2024). We explore whether those improvements are reflected in discourse performance.

MAP. A MT model defines a probability distribution $p(y|x, \theta)$ over a set of hypotheses $\mathcal{Y}$. MAP decoding, such as greedy decoding, aims to maximize the probability of generated hypothesis:

$$\hat{h} = \arg\max_{y \in \mathcal{Y}} p(y|x, \theta). \tag{1}$$

MBR. Given a utility function u that measures the similarity between a hypothesis h and a reference y, MBR decoding aims to find the hypothesis that maximizes the expected utility (minimizes the loss) among a set of sampled hypotheses $\mathcal{H}$ (each sampled hypothesis is used as a pseudo reference and compared to all other hypotheses). It selects:

$$\hat{h} = \arg\max_{h \in \mathcal{H}} \mathbb{E}_{y \sim p(y|x,\theta)} \left[u(h, y) \right]. \tag{2}$$

We experiment with different choices of utility functions, including lexical, pretrained, and discourse-specific metrics for translation evaluation. Additionally, we experiment with other inference approaches, including fusion and automatic post-editing. We discuss these in detail in §5.1.

3 Experiments[2]

3.1 Data

We experiment on the DELA corpus (Castilho et al., 2021), an English-Brazilian-Portuguese document-level corpus annotated with context-related issues. The corpus is a collection of 60 documents with 3.7K sentences from different domains (news, subtitles, literature, legislation, reviews, medical) that are manually selected, translated, and annotated with context-dependent discourse phenomena. Additionally, we experiment on a subset of TED2020 data (20K sentences in around 160 documents) (Reimers and Gurevych, 2020) available in OPUS (Tiedemann, 2012). We also experiment on the WMT24++ dataset (Deutsch et al., 2025), which has a size of approximately 1K sentences across 169 documents (results are in Appendix B). For both TED2020 and WMT24++, we experiment on six language directions: English (EN) to Brazilian-Portuguese (PT), German (DE), French (FR), Korean (KO), Arabic (AR) and Russian (RU). Dataset statistics for the three corpora, including discourse phenomena statistics, are presented in Appendix A.

3.2 Models

We prioritize high performing models in multilingual and MT-related tasks. We evaluate TowerInstruct-13B (Alves et al., 2024), an instruction-tuned translation-LLM based on Llama2-13B (Touvron et al., 2023), which is a leading translation system according to WMT24 (Kocmi et al., 2024). We chose the best-performing version of the model within our budget. We also evaluate EuroLLM-9B-Inst (Martins et al., 2024), a leading European LLM[3] trained from scratch on all European Union languages and additional relevant ones. EuroLLM-9B-Inst is trained on all translation data used for TowerInstruct-13B, as well as additional sources, giving it broader coverage.

For generalizability, we include LLMs with strong performance on general language generation tasks, namely Gemma3-12B-it (Kamath et al., 2025) and Qwen3-14B (Yang et al., 2025). As an encoder-decoder baseline, we include NLLB-3.3B (Costa-jussà et al., 2022), a leading sentence-level multilingual NMT model with strong performance across languages of interest in this work. Context-

[2] Sustainability statement for all experiments: Appendix H.

[3] based on the european-llm-leaderboard https://huggingface.co/spaces/Eurolingua/european-llm-leaderboard

	BLEU	COMET	docCOMET	COMETQE	docCOMETQE	L.repetition	Formality	Pronouns
ctx= 0								
nllb G	55.2	87.1	82.2	81.5	75.4	**87.0**	**75.0**	45.0
nllb Q	**58.2**	87.4	82.3	81.6	76.5	85.0 ↓	**76.0**	47.0
tower G	34.4	84.7	79.1	79.7	74.6	77.0	51.0	32.0
tower Q	52.7	88.9	84.3	**83.0**	**80.1**	85.0	68.0	39.0
euro G	30.7	82.5	75.2	77.9	70.1	78.0	59.0	33.0
euro Q	52.4	87.9	81.5	82.5	79.8	86.0	**75.0**	45.0
gemma G	50.6	85.6	80.1	79.4	75.9	82.0	12.0	38.0
gemma Q	53.0	88.5	83.5	82.3	78.8	84.0	72.0	41.0
qwen G	50.9	88.5	83.6	82.1	78.9	84.0	**73.0**	41.0
qwen Q	54.1	88.9	84.0	82.4	79.3	84.0	**74.0**	40.0 ↓
ctx= 5								
tower G	41.0	86.0	81.2	79.1	74.5	85.0	66.0	50.0
tower Q	57.4	**89.6**	**85.4**	82.0	79.5	**90.0**	**76.0**	**60.0**
euro G	25.9	80.4	73.6	76.0	66.1	79.0	56.0	40.0
euro Q	52.1	87.8	82.0	81.9	78.9	**89.0**	**75.0**	48.0
gemma G	54.0	87.3	82.3	80.3	77.4	86.0	27.0	**49.0**
gemma Q	56.2	88.5	83.8	81.5	78.5	87.0	51.0	**52.0**
qwen G	52.4	89.3	84.9	82.0	79.3	**89.0**	**75.0**	**51.0**
qwen Q	52.4	**89.6**	**85.3**	82.2	79.7	**90.0**	**76.0**	**52.0**

Table 2: Translation and discourse phenomena performance of all models (nllb=NLLB-3.3B, tower=TowerInstruct-13B, euro=EuroLLM-9B-Inst, gemma=Gemma3-12B-it, qwen=Qwen3-14B) using greedy (G) and quality-aware decoding (Q) on **DELA dataset** in both sentence-level (ctx= 0) and context-aware (ctx= 5) setups. The encoder–decoder baseline (nllb=NLLB-3.3B) is highlighted in gray. **Bold** highlights the best value per column (with statistical significance p<0.05). The numbers presented for phenomena are F1 accuracies (details in §3.4.2). Cases where the QAD result is worse than its Greedy counterpart are marked with a red down arrow ↓. Cases where the context-aware result is better than the sentence level result are marked with a green underline.

aware encoder-decoder variants (see §6) would also be worth exploring, but have limited language coverage and are beyond the scope of this work.

3.3 Inference

We compare two decoding setups: **greedy** decoding,[4] which selects the highest-probability token at each step, and quality-aware decoding (**QAD**), which uses MBR with 50 samples generated via nucleus sampling (p=0.9).[5] We use **BLEU** score (Papineni et al., 2002) as utility function for all our experiments unless otherwise indicated. Note that we first conducted preliminary experiments on different prompting formats for each model and present only the best setup in this work. For LLMs (all models except NLLB-3.3B), we employ context-aware prompting with the context being (up to) 5 previous source-target pairs in the same document (prompt formats in Appendix E). We opt for the context-aware setup instead of translating documents holistically to eliminate the potential bias of how models handle different parts of the context (Liu et al., 2024). For NLLB-3.3B, since the model has only been trained on sentence-level data, we conduct inference at the sentence level. We also report sentence-level baseline results for LLMs.

3.4 Evaluation

3.4.1 Overall Translation Evaluation

We use a lexical metric, BLEU[6] (Papineni et al., 2002), a reference-based pretrained metric, COMET[7] and its document-level variant docCOMET (Rei et al., 2022a), and a reference-free pretrained metric, COMETQE[8] (Rei et al., 2022b) and its document-level variant docCOMETQE.

3.4.2 Discourse Phenomena Evaluation

We measure F1 accuracy of tagged words with discourse phenomena in the reference being tagged in the hypothesis. To do so, we utilize the multilingual discourse-aware benchmark (MuDA) for discourse phenomena evaluation (Fernandes et al., 2023). Tagging for words that require inter-sentential context is done automatically using predefined language-specific lists of pronouns, verb forms, and formality indicators. For lexical repetition, tag-

[4] Beam search decoding for NLLB yielded comparable results to greedy decoding. Details in Appendix K.

[5] For Qwen3-14B, we use the "non-thinking" mode with the sampling parameters recommended by its developers: t=0.7, topP=0.8, topK=20, minP=0. https://huggingface.co/Qwen/Qwen3-14B

[6] SacreBLEU signatures in Appendix I

[7] https://huggingface.co/Unbabel/wmt22-comet-da

[8] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

	BLEU	COMET	docCOMET	COMETQE	docCOMETQE	L.repetition	Formality	Pronouns
ctx= 0								
nllb G	28.3	84.2	79.2	82.7	80.0	64.2	57.4	61.2
nllb Q	29.3	84.3	79.4	82.6 ↓	80.0	64.7	57.6	60.5 ↓
tower G	21.0	81.5	76.1	80.1	75.1	60.0	48.8	51.3
tower Q	31.1	85.7	80.9	**84.2**	**81.9**	66.4	58.2	60.3
euro G	15.8	79.6	73.0	78.3	71.0	56.8	48.0	50.0
euro Q	26.6	84.7	79.0	83.5	81.1	63.8	59.0	59.0
gemma G	25.1	81.9	76.2	80.5	78.3	60.3	16.4	55.2
gemma Q	26.4	84.4	79.1	83.2	80.7	62.5	55.4	57.5
qwen G	24.8	83.4	78.1	82.2	79.0	61.8	50.4	57.8
qwen Q	25.9	84.4	79.1	83.2	80.8	63.3	52.2	59.0
ctx= 5								
tower G	21.0	79.6	74.8	75.1	71.4	62.0	52.2	62.0
tower Q	**33.0**	**86.1**	**82.0**	82.5	80.1	**71.2**	**64.6**	71.0
euro G	15.0	78.7	72.7	77.0	68.9	60.5	49.4	52.8
euro Q	28.2	85.1	79.9	83.2	80.8	67.7	62.4	63.0
gemma G	26.9	80.9	75.4	79.1	76.6	63.2	18.4	58.8
gemma Q	27.8	83.8	78.5	82.3	80.0	64.8	42.6	58.8
qwen G	26.7	84.4	79.4	82.3	79.5	66.7	61.0	61.2
qwen Q	27.9	85.2	80.4	83.2	81.1	67.0	62.0	**82.8**

Table 3: Translation and discourse phenomena performance of all models (nllb=NLLB-3.3B, tower=TowerInstruct-13B, euro=EuroLLM-9B-Inst, gemma=Gemma3-12B-it, qwen=Qwen3-14B) using greedy (G) and quality-aware decoding (Q) on **TED2020 dataset** in both sentence-level (ctx= 0) and context-aware (ctx= 5) setups. The results are averaged across all language pairs. The encoder–decoder baseline (nllb=NLLB-3.3B) is highlighted in gray. **Bold** highlights the best value per column. Cases where the QAD result is worse than its Greedy counterpart are marked with a red down arrow ↓. Cases where the context-aware result is better than the sentence level result are marked with a green underline.

ging is performed by obtaining source-target word alignments; if an alignment pair occurs more than a specific number of times in the document (three in our experiments, following MuDA), the word is tagged for lexical repetition.

3.4.3 LLM-Based Evaluation

Evaluating LLMs automatically has become increasingly difficult due to their rapid advancements. Consequently, using language models to automatically assess long-form text (LLM-as-a-judge) is gaining popularity. We employ the multilingual M-Prometheus (Pombal et al., 2025) judge in an absolute evaluation setup where the judge is provided with the instruction used to prompt the translation model, along with the translation output. The judge assigns a rating between 1 and 5, accompanied by an explanation of the decision. Since we use different prompting setups (§3.3), the instructions provided to the judge are different, which makes direct comparisons unfair. Therefore, we report only the difference between greedy and QAD scores for each model rather than their absolute scores.

4 Results

In Table 2, we present the results on the DELA corpus. We see that LLMs often fall behind NLLB-3.3B in translation and discourse phenomena performance when using greedy decoding. Interestingly, we observe that using QAD notably improves both overall translation and discourse phenomena handling of LLMs, allowing them to outperform NLLB-3.3B.[9] In Table 3 we show the results on the TED2020 dataset averaged across all language pairs. WMT24++ results are deferred to Appendix B as they evidence similar overall trends. Detailed language-specific results are in Appendix C. The results highlight the substantial improvements in discourse and translation performance of translation-finetuned LLMs (TowerInstruct-13B and EuroLLM-9B-Inst) using QAD. The performance of general-purpose LLMs (Qwen3-14B and Gemma3-12B-it) is also improved, though the gains are more modest, which could be attributed to the lack of translation-specific specialization in those models. Among all languages and datasets tested, TowerInstruct-13B achieves the best overall performance, highlighting the effectiveness of specialized translation finetuning in encoding discourse awareness in LLMs. Additionally, we present the differences

[9] Although NLLB performs well on discourse phenomena, without access to context it guesses outputs that may align with intended translations but remain inconsistent and unprincipled.

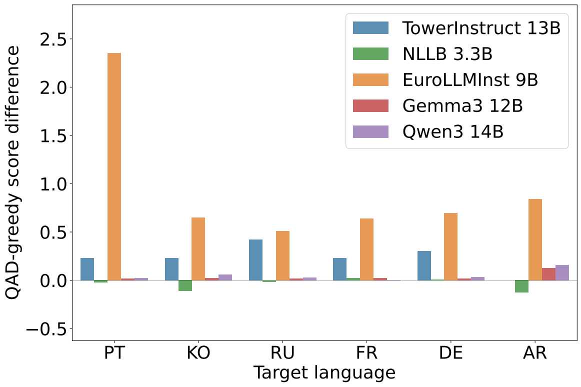


Figure 1: Difference between QAD and greedy LLM-as-a-judge scores on TED2020 data. The top-to-bottom order of the legend corresponds to the left-to-right order of the bars in each language group. The plot demonstrates that QAD improves the performance of LLMs.

between greedy and QAD scores from the LLM-as-a-judge evaluation for TED2020 data in Figure 1, WMT24++ results are in Appendix B. The results, consistent with automatic metrics, show that QAD noticeably enhances the performance of (TowerInstruct-13B, EuroLLM-9B-Inst), while the improvements for Gemma3-12B-it and Qwen3-14B are relatively modest. In contrast, QAD does not improve on the performance of NLLB-3.3B.

From Tables 2 and 3, the effectiveness of QAD is evident in both sentence-level and context-aware setups of LLMs. Although sentence-level setups sometimes achieve higher overall translation performance (specially with greedy decoding), context-aware setups consistently achieve better performance on discourse phenomena. Notably, COMETQE and docCOMETQE tend to rate sentence-level outputs higher than context-aware ones, unlike other metrics. Through further analysis and manual inspection, we find that while fluency and correctness are comparable, sentence-level outputs are shorter on average, suggesting a bias in COMETQE toward brevity. Similar bias toward shorter tanslations has been observed for other COMET variants (e.g. XCOMET; Guerreiro et al. (2024)) in Perrella et al. (2024). In combination with our findings, we hope it motivates further research on the behavior and robustness of such metrics for context-aware translation evaluation.

5 Analysis

5.1 Inference Setup Ablation

We perform an ablation study on the entire DELA data using the TowerInstruct-13B model, comparing different inference setups. Specifically:

QAD. We explore the following utility functions:

- **Translation metrics.** BLEU, ChrF (Popović, 2015) and COMET scores. Here, we use MBR with 50 samples-per-instance generated using nucleus sampling (p=0.9).
- **Discourse-specific metrics.** Lexical cohesion (LC) ratio (Wong and Kit, 2012), which is the number of lexical cohesion devices (repetitions, hypernyms, and synonyms) divided by the total number of content words, and DiscoScore (Zhao et al., 2023), a parametrized metric that uses BERT (Devlin et al., 2019) to model discourse coherence through sentence graphs. Here, we use MBR with 20 samples-per-instance generated using nucleus sampling (p=0.9).[10]

Fusion. Proposed by Vernikos and Popescu-Belis (2024), the approach works by combining spans from different candidates generated via nucleus sampling (p=0.9) using a QE metric (COMETQE).

Automatic post editing (APE). Editing greedy outputs using XTOWER (Treviso et al., 2024) and XCOMET (Guerreiro et al., 2024), as in the IT-Unbabel team's submission to the quality estimation shared task at WMT24 (Zerva et al., 2024).

We assess the methods based on translation and discourse phenomena performance. Based on Table 4, QAD outperforms other inference approaches, including fusion and APE. Notably, translation metrics are more effective utility functions compared to discourse-specific metrics, with lexical measures (BLEU, ChrF) slightly outperforming the pretrained COMET, though overall performance remains comparable. Statistical significance testing, using greedy decoding as the baseline, shows significant improvements across all metrics for QAD with BLEU, ChrF, COMET, and Disco. However, QAD (LC), fusion, and APE are not outperforming the greedy baseline for discourse phenomena. In Appendix F, we report the runtime details for all inference setups, demonstrating that QAD with translation metrics (specifically BLEUand ChrF) offers the best trade-off between translation quality and computational cost.

Inspired by Zerva et al. (2024), to gain deeper insight into the discourse-related revisions intro-

[10] We use 20 samples instead of 50 due to computational constraints, as the discourse metrics involve generating an entity graph for each sample, which becomes impractical with a higher number of samples.

	BLEU	ChrF	COMET	docCOMET	COMETQE	docCOMETQE	Rep.	Formal.	Pro.
Greedy	41.0	66.6	86.0	81.2	79.1	74.5	85	66	50
QAD(BLEU)	**57.4***	76.3*	89.6*	85.4*	82.0*	79.5*	**90***	**76***	**60***
QAD(ChrF)	55.8*	**76.9***	89.7*	85.4*	82.2*	79.6*	**90***	**77***	**61***
QAD(COMET)	54.2*	75.0*	**90.9***	**86.2***	83.1*	80.7*	**89***	**76***	**62***
QAD(LC)	41.3	67.5*	84.6*	81.8*	79.6*	75.5*	85	64	49
QAD(Disco)	55.3*	75.1*	89.4*	85.0*	81.8*	78.9*	**89***	**76***	**57***
Fusion	41.6	67.8*	89.1*	84.1*	**85.7***	**82.5***	86	67	46
APE	44.3*	68.4*	87.7*	82.6*	82.0*	77.9*	84	69	47

Table 4: Translation and discourse phenomena performance of different decoding setups using TowerInstruct-13B on DELA data. **Bold** highlights the best value per column (with statistical significance p<0.05). Statistically significant outputs compared to the greedy output are marked with "*" (paired bootstrap resampling (Koehn, 2004)).

Setup	Edit rate
QAD (BLEU)	32.7
QAD (ChrF)	33.6
QAD (COMET)	35.4
QAD (LC)	46.4
QAD (Disco)	34.5
Fusion	45.1
APE	29.6

Table 5: Edit rate against the greedy output.

duced by each setup, we analyze the edit rate in their outputs relative to greedy decoding, using it as a *proxy* for discourse-related revisions. The analysis focuses on sentences tagged with discourse phenomena using MuDA (Fernandes et al., 2023). In Table 5, we present the edit rate of each setup compared to the greedy output. Tables 4 and 5 show that the edit rate is aligned with discourse and translation performance. A moderate level of edit operations (insert, delete, substitute, shift) produces strong translation and discourse performance results, as demonstrated by the utility functions BLEU, ChrF, COMET, and DiscoScore. However, deviations from this balance, whether through fewer edits (APE) or excessive edits (LC and fusion), compromise performance. Overall, this analysis highlights that among the experimented setups, **QAD with translation metrics is the best setup to improve discourse performance**.

5.2 Human Evaluation and Qualitative Analysis

We conduct a small-scale manual evaluation of the outputs to examine the semantic differences between greedy and QAD outputs, and to confirm the findings of the automated evaluation of discourse phenomena, which relies on MuDA (Fernandes et al., 2023). We use the outputs of the best-performing model on TED2020 data, which is TowerInstruct-13B for all languages except Arabic, where we use EuroLLM-9B-Inst. We randomly sample a subset of 25 instances of {`source`, `greedy_MT`, `QAD_MT`} for each language, all annotated with discourse phenomena via MuDA (Fernandes et al., 2023) and accompanied with preceding context. We provide these to native or bilingual speakers —who voluntarily participated in the annotation process— as we are interested in how non-expert translators from the general public perceive

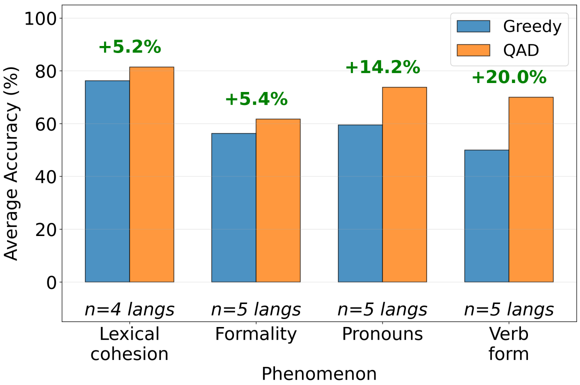


Figure 2: Human-annotated accuracy of greedy and QAD outputs in handling discourse phenomena, averaged across all languages where the phenomena occur (number of languages is at the bottom). Arabic is excluded from this plot to avoid model-specific biases.

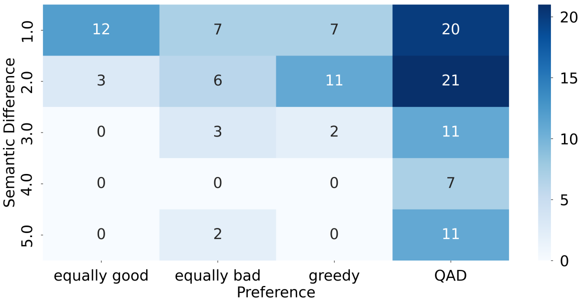


Figure 3: Semantic difference vs. preference summed over all languages (except Arabic).

the translations (discourse-related quality improvements should be perceivable by the general population, as they are highly relevant to communication and conversation quality). We mask the MT type information[11] and ask them to annotate the data as follows: first, identify any of the four linguistic phenomena present in the source sentence. Once identified, they determine whether the phenomenon is translated correctly in (a) the greedy translation and (b) the QAD translation. Next, they annotate the semantic difference between the greedy and QAD hypotheses using a Likert scale ranging from 1 to 5. They then select their preferred translation between the two hypotheses and may optionally provide comments explaining their preference and observations (full guidelines in Appendix G).

Annotation statistics are presented in Appendix A, where we see a high correlation between automatic tags with MuDA and human tags (with an overlap of 60%-100%). Figure 2 presents the average performance of greedy and QAD outputs across languages, showing improved performance for QAD across all phenomena, which aligns with the results of the automatic evaluation.

Results of the semantic difference against preferences are presented in Figure 3 (Arabic is excluded from these figures to remove model-specific bias; its results in Appendix D confirm the same findings). We notice that QAD output is generally preferred, while greedy output tends to be less frequently chosen as the preferred option, especially when there are larger semantic differences between the outputs. Greedy output is still sometimes preferred in cases where the semantic differences are smaller. These patterns suggest that **QAD generates semantically richer samples that align with human preferences compared to greedy decoding**. In addition, analyzing the comments we received from the participants, it seems that QAD-based outputs are closer to human perception in terms of discourse and fluency, even when translation errors occur.

Finally, in Appendix J, we present the details and findings of a qualitative analysis on TED2020 data. We analyze sentence-level and context-aware greedy and QAD outputs and showcase specific examples demonstrating how the context-aware QAD setup unlocks models' ability to handle different discourse phenomena such as pronoun disambiguation, lexical repetition, formality and verb forms.

6 Related Work

We provide an overview of research on document-level MT, covering model architectures, adaptation of LLMs for document-level translation, and evaluation metrics for assessing discourse performance.

Document-level translation. Document-level translation models allow incorporating inter-sentential dependencies during translation. Architectures include single-encoder approaches (Tiedemann and Scherrer, 2017; Bawden et al., 2018), multi-encoder approaches (Libovický and Helcl, 2017; Zoph and Knight, 2016; Wang et al., 2017; Bawden et al., 2018; Zhang et al., 2018; Li et al., 2020; Lupo et al., 2022), multi-hop transformers (Zhang et al., 2021), hierarchical context-encoders (Wang et al., 2017; Miculicich et al., 2018; Maruf et al., 2019; Yun et al., 2020), and document-embedding approaches (Huo et al., 2020; Macé and Servan, 2019; Morishita et al., 2021).

LLMs for document-level translation. LLMs are becoming increasingly competitive in MT (Kocmi et al., 2024). Efforts to adapt LLMs for document-level MT include finetuning models using mixed sentence-level and document-level instructions (Li et al., 2024b; Ramos et al., 2025), prompting models via in-context learning (Cui et al., 2024), hybrid techniques combining sentence-level translation models and monolingual document-level language models (Petrick et al., 2023), and agentic systems based on multi-level memory (Wang et al., 2025).

Discourse evaluation in machine translation. Studying discourse characteristics in MT has been going on since the early works on rule-based and statistical MT systems (Hardmeier, 2012) and continued as NMT models prevailed (Tan et al., 2022a; Honda et al., 2023; Luo et al., 2024). Subsequent studies have developed discourse-specific datasets and benchmarks, including test sets for coreference, entities, terminology, quotations, and readability (Jiang et al., 2023; Jwalapuram et al., 2020; Müller et al., 2018; Lopes et al., 2020). Research efforts have also focused on designing evaluation metrics to assess discourse performance, such as cohesion, coreference resolution, terminology consistency, and zero pronoun translation (Tan et al., 2022b; Bawden et al., 2018; Wu et al., 2024a; Wang et al., 2023).

[11] Annotators see the translation hypotheses as pairs of `output_1, output_2`

7 Conclusion

We investigate the discourse phenomena performance of LLMs in context-aware translation. Specifically, we examine four aspects of discourse: lexical repetition, formality, pronoun resolution, and verb forms. Our findings reveal that LLMs still exhibit weaknesses in discourse performance when using greedy decoding. To address this limitation, we propose the use of quality-aware decoding (QAD) to better leverage the discourse knowledge encoded within LLMs. We demonstrate the effectiveness of QAD through extensive automatic evaluations across six language pairs and two datasets. Additionally, we conduct an ablation study comparing different decoding methods and perform a human assessment on a subset of the data to analyze the lexical and semantic changes introduced by QAD. To support further research, we release the dataset with human annotations of discourse phenomena. Future research can explore the use of such annotated data as a reward signal for fine-tuning LLMs to further enhance their discourse phenomena performance.

Limitations

- We use MuDA (Fernandes et al., 2023) as it provides an automatic method for tagging various discourse phenomena. However, it does not capture all aspects of discourse in translation such as zero pronoun translation, readability, etc. Additionally, we adopt MuDA's default alignment and coreference resolution models, which may not reflect the current state of the art. Enhancing these components and employing a more comprehensive discourse evaluation are directions for future research.

- We experiment with only one sampling approach and one hyperparameter setup for it (nucleus sampling with p=0.9). We leave it to future research to investigate the effect of the number of samples, the sampling method used and its hyper-parameters on discourse performance.

- We perform the LLM-as-a-judge evaluation at the overall translation level, as we utilize an off-the-shelf model that was not sensitive to specific phenomena changes. Future work could focus on adapting LLM judges to discourse phenomena evaluation.

- We attempt to cover as many languages and models as possible, given the experimental resources we have. Additional observations may arise for languages and models we did not cover.

- We perform the human evaluation on a limited amount of data. Based on our conclusions, it would be useful to have a larger dataset with human annotations, which would allow for more comprehensive experiments, supervision of models, etc.

- The performance on the discourse phenomena we study can be affected by gender bias in the tested models, we leave it to future work to study the correlation between models' bias and discourse performance.

Ethical Considerations

Machine translation is a widely adopted technology, sometimes in sensitive, high-risk settings. Even though we perform a thorough analysis of LLMs' performance on discourse phenomena during translation, and propose the use of quality aware-decoding to improve the performance, we still rely heavily on automatic evaluation which is imperfect. For systems deployed in critical scenarios, we advocate for detailed, case-specific assessments to ensure reliability.

Acknowledgements

We would like to thank Hippolyte Gisserot-Boukhlef, Sergey Troshin, Beatriz Canaverde and Emmanouil Zaranis for their valuable discussions about this work, and their help with the annotations. This work is part of the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631. WM acknowledges travel support from ELIAS (GA no 101120237). VN was partially supported by the Dutch Research Council (NWO) via VI.Veni.212.228. CZ was partially funded by the FCT project "OptiGov" (DOI 10.54499/2024.07385.IACDC), funded by the PRR under the measure RE-C05-i08.m04. She is also supported by the PRR Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações. Finally, she is supported by an unrestricted gift from Google (Google Research Scholar).

References

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. 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. *CoRR*, abs/2402.17733.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. 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 4528–4537, Online. Association for Computational Linguistics.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. DELA corpus - a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 566–577, Online. Association for Computational Linguistics.

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, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level. Technical report, Inria Paris, Sorbonne Université ; Sorbonne Universite ; Inria Paris.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: expanding the language coverage of WMT24 to 55 languages & dialects. *CoRR*, abs/2502.12404.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

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

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sumire Honda, Patrick Fernandes, and Chrysoula Zerva. 2023. Context-aware neural machine translation for english-japanese business scene dialogues. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 272–285.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq R. Joty, and Dominika Basaj. 2020. Can your context-aware MT system pass the dip benchmark tests? : Evaluation benchmarks for discourse phenomena in machine translation. *CoRR*, abs/2004.14607.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 191 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference, CI 2023, Delft, Netherlands, November 6-9, 2023*, pages 12–24. ACM.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural 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 7250–7264, Online. Association for Computational Linguistics.

Baohang Li, Zekai Ye, Yichong Huang, Xiaocheng Feng, and Bing Qin. 2024a. SCIR-MT‘s submission for WMT24 general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 280–285, Miami, Florida, USA. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024b. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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.

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li, Jinlong Yang, and Hao Yang. 2024. Context-aware and style-related incremental decoding framework for discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 973–979.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Valentin Macé and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *CoRR*, abs/2409.16235.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2024a. Analyzing context utilization of llms in document-level translation. *arXiv preprint arXiv:2410.14391*.

Wafaa Mohammed and Vlad Niculae. 2024b. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta. Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, Tomoharu Iwata, and Masaaki Nagata. 2021. Context-aware neural machine translation with mini-batch embedding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2513–2521, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural 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 259–272, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 375–391, Singapore. Association for Computational Linguistics.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *Preprint*, arXiv:2504.04953.

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.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André F. T. Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. *CoRR*, abs/2504.12140.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. 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. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022a. Towards discourse-aware document-level neural machine translation. In *IJCAI*, pages 4383–4389.

Xin Tan, Longyin Zhang, and Guodong Zhou. 2022b. Discourse cohesion evaluation for document-level neural machine translation. *CoRR*, abs/2208.09118.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the*

Association for Computational Linguistics: EMNLP 2024, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

Giorgos Vernikos and Andrei Popescu-Belis. 2024. Don‘t rank, combine! combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.

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

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.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dawid Wisniewski, Zofia Rostek, and Artur Nowakowski. 2024. FAME-MT dataset: Formality awareness made easy for machine translation purposes. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 164–180, Sheffield, UK. European Association for Machine Translation (EAMT).

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024b. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC‘s submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 155–164, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. 2020. Improving context-aware neural machine translation using self-attentive sentence embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9498–9506. AAAI Press.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Long Zhang, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *CoRR*, abs/2403.00277.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association* *for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

A Dataset statistics

Table 9 presents the dataset statistics for the three corpora (DELA, TED2020, WMT24++), including discourse phenomena statistics as well as the number of sentences and documents. Table 6 presents the annotation statistics of discourse phenomena in the human assessment analysis.

B WMT24++ Results

Table 10 shows the results on WMT24 ++ for averaged across all language pairs. LLM-as-a-judge scores are shown in Figure 4.

C Detailed Language-Specific Results

Detailed language-specific results for all models on TED2020 and WMT24++ datasets are shown in Tables 11 and 12, respectively. It is worth mentioning that all models exhibit low performance on EN-DE WMT24++ data. A manual qualitative analysis of the translations reveals that the reference translations are of suboptimal quality, often consisting of short sentences.

D Results of Human Qualitative Analysis on Arabic

Figure 5 shows the human annotations of the performance of QAD and greedy outputs on Arabic. Figure 6 shows the preference and semantic difference relationship for Arabic.

E Prompt Formats

Figures 7 to 9 present the prompt formats.

	rep.	Formal.	Pro.	Verb.	Total	(%)
EN-PT	12	1	7	1	15	60
EN-DE	16	5	15	6	22	88
EN-FR	5	14	10	9	21	84
EN-KO	0	2	4	25	25	100
EN-AR	6	0	10	2	16	64
EN-RU	4	7	7	6	16	64

Table 6: Human-annotated discourse phenomena statistics, including counts of each phenomenon, the total number of sentences tagged with phenomena and their percentage of total sentences. Note that the total column can be less than the sum of phenomena columns because we can have multiple phenomena per sentence.

Setup	Runtime
Greedy	1.1 min
QAD (BLEU)	43.7 min
QAD (ChrF)	55.3 min
QAD (COMET)	11.7 hr
QAD (LC)	17 hr
QAD (Disco)	41.3 hr
Fusion	35.2 min
APE	5.7 min

Table 7: Runtime details for all inference setups.

F Runtime Details

In Table 7 we present the runtime details of all inference setups tested in §5.1. All experiments are carried out under identical computational conditions: a single H100 GPU with 192 GB of memory.

G Human Assessment Details

Details of the data and instructions given to the annotators are presented in Table 14.

H Sustainability Statement

Our experiments run in 782h on 1 GPU NVIDIA A100 40GB PCIe, and draw 334.06 kWh. Based in [redacted for anonymity], this has a carbon footprint of 125.05 kg CO2e, which is equivalent to 11.37 tree-years (Lannelongue et al., 2021).

I SacreBLEU Signatures

To ensure reproducability, we present SacreBLEU signatures for BLEU and ChrF metrics in Table 8.

J Qualitative Examples

With the help of human annotators, we manually analyze the outputs of the sentence-level and context-aware greedy and QAD setups using EuroLLM-

metric	signature
BLEU	nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1
ChrF	nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1

Table 8: Evaluation-metrics signatures

9B-Inst model for Arabic and TowerInstruct-13B model for other languages to understand how QAD improves specific discourse phenomena. The analysis is performed on TED2020 dataset. We summarize our findings and present specific examples:

- **Dual-pronoun disambiguation:** using context-aware QAD, the reference of the pronoun *they* to a dual subject in Arabic is correctly resolved, and the appropriate form is applied (Figure 10).
- **Plural-pronoun disambiguation:** using context-aware QAD, the reference of the pronoun *you* to a plural subject in Arabic is correctly resolved, and the appropriate form is applied (Figure 11).
- **Lexical repetition:** using context-aware QAD, the entity *Lakota* is translated consistently with the context in Arabic (Figure 12).
- **Plural-pronoun and past verb form:** using context-aware QAD, the reference of the pronoun *you* to a plural subject in Russian is correctly resolved. Additionally, the correct past verb form *said* is used. Context-aware QAD is the only setup where both phenomena are handled (Figure 13).
- **Formality and pronoun:** using context-aware QAD, the correct form of the singular second person pronoun *you* in Brazilian-Portuguese is used, it also maintains the correct formality level of explicitly mentioning the pronoun which conveys a conversational tone rather than an implicit pronoun which is more formal and detached (Figure 14).

K Beam Search Decoding For NLLB

Prior studies have shown the effectiveness of beam search decoding for encoder–decoder models (Sutskever et al., 2014; Freitag and Al-Onaizan, 2017). In Table 13, we report the results of greedy and beam search decoding for NLLB-3.3B on the DELA dataset. We observe comparable performance across the two decoding strategies and therefore adopt greedy decoding to ensure consistency and comparability with the other models.

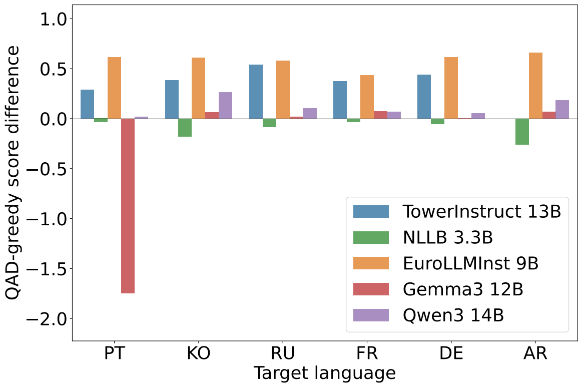


Figure 4: Difference between QAD and greedy LLM-as-a-judge scores on WMT24++ data. The top-to-bottom order of the legend corresponds to the left-to-right order of the bars for each language. The plot demonstrates that QAD improves the performance of LLMs.

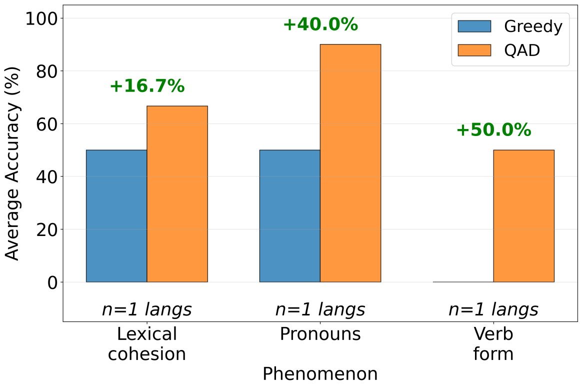


Figure 5: Human-annotated accuracy of greedy and QAD outputs in handling discourse phenomena for Arabic data.

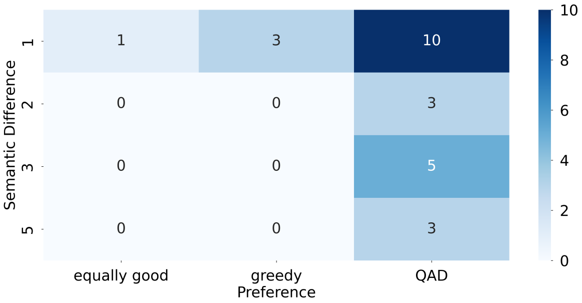


Figure 6: Semantic difference vs. preference on Arabic data.

	Lexical repetition	Formality	Pronouns	Verb form	Total	Sentences	Documents
DELA							
EN-PT	1322	630	323	–	1866 (50.3)	3710	60
TED2020							
EN-PT	6640	3151	2202	–	9877 (49.4)	20003	162
EN-DE	5386	4904	2186	–	10125 (50.4)	20077	160
EN-FR	6346	3315	7486	–	11642 (58.1)	20049	162
EN-KO	2190	1165	–	–	3238 (16.2)	20017	162
EN-AR	4109	–	655	–	4654 (23.2)	20034	162
EN-RU	3544	2451	–	–	5506 (27.4)	20084	163
WMT24							
EN-PT	209	178	59	–	356 (37.1)	960	169
EN-DE	56	199	43	–	263 (27.4)	960	169
EN-FR	189	130	160	67	413 (43.0)	960	169
EN-KO	93	17	–	–	109 (11.4)	960	169
EN-AR	166	–	39	–	198 (20.6)	960	169
EN-RU	129	90	–	70	255 (26.6)	960	169

Table 9: Dataset statistics, including counts of each phenomenon, the total number of sentences tagged with phenomena and their percentage of total sentences (in parentheses), and the total number of sentences and documents for each dataset and language pair. Note that the Total column can be less than the sum of phenomena columns because we can have multiple phenomena per sentence.

	BLEU	COMET	docCOMET	COMETQE	docCOMETQE	L.repetition	Formality	Pronouns	Verb form
ctx= 0									
nllb G	20.3	72.5	70.3	75.8	67.4	56.5	40.8	44.2	34.5
nllb Q	22.2	73.3	71.0	76.5	68.6	53.3 ↓	43.6	42.2 ↓	41.5
tower G	16.9	72.3	69.9	74.8	66.3	49.8	33.8	28.7	26.5
tower Q	24.8	76.2	73.8	**80.6**	**73.8**	56.6	43.4	36.7	39.5
euro G	14.1	67.9	65.1	72.9	63.0	45.7	36.4	36.8	28.5
euro Q	21.5	72.0	69.2	79.6	72.6	57.5	45.4	41.2	**43.0**
gemma G	21.2	71.7	68.9	78.0	71.4	52.8	44.2	41.0	41.0
gemma Q	22.3	73.0	70.2	79.7	72.9	55.3	48.4	46.0	**43.0**
qwen G	19.5	71.4	68.8	78.2	70.8	52.3	41.2	44.0	36.5
qwen Q	20.8	72.3	69.6	79.8	72.9	54.7	40.0 ↓	43.2 ↓	39.5
ctx= 5									
tower G	17.0	73.0	71.5	69.0	63.5	54.2	39.8	38.0	25.5
tower Q	**25.7**	**78.5**	**76.7**	76.0	72.0	60.2	**49.2**	41.0	41.5
euro G	15.9	71.9	69.6	71.6	63.2	55.8	37.8	38.8	32.5
euro Q	24.9	77.1	74.8	78.5	72.1	**65.3**	46.6	46.5	**43.0**
gemma G	22.1	71.0	68.4	75.7	70.4	61.5	43.6	42.2	38.5
gemma Q	23.4	72.6	70.1	78.5	72.5	64.7	50.8	43.0	42.5
qwen G	20.5	72.1	69.6	78.2	71.1	62.5	47.6	44.8	38.5
qwen Q	21.7	72.8	70.4	79.6	73.0	64.2	42.2 ↓	**48.0**	39.0

Table 10: Translation and discourse phenomena performance of all models (nllb=NLLB-3.3B, tower=TowerInstruct-13B, euro=EuroLLM-9B-Inst, gemma=Gemma3-12B-it, qwen=Qwen3-14B) using greedy (G) and quality-aware decoding (Q) on **WMT24++ dataset** in both sentence-level (ctx= 0) and context-aware (ctx= 5) setups. The results are averaged across all language pairs. The encoder–decoder baseline (nllb=NLLB-3.3B) is highlighted in gray. **Bold** highlights the best value per column. Cases where the QAD result is worse than its Greedy counterpart are marked with a red down arrow ↓. Cases where the context-aware result is better than the sentence level result are marked with a green underline.

```
Translate the following <src_lang> source text to <tgt_lang>.
<src_lang>: <src context 1> <src context 2> <src context 3> <src context 4> <src context 5>
↪  <src_sentence>
<tgt_lang>: <tgt context 1> <tgt context 2> <tgt context 3> <tgt context 4> <tgt context 5>
```

Figure 7: TowerInstruct-13B prompt format.

	Metric	nllb		tower				euro				gemma				qwen			
		ctx= 0		ctx= 0		ctx= 5		ctx= 0		ctx= 5		ctx= 0		ctx= 5		ctx= 0		ctx= 5	
		G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q
EN-PT	BLEU	40.4	41.8	26.5	38.5	30.4	**42.5**	24.7	37.4	21.0	38.9	38.1	39.2	40.2	39.1	38.3	39.2	40.1	41.0
	COMET	87.0	87.2	83.6	87.5	84.6	**88.2**	82.7	86.8	81.1	87.2	85.0	87.1	85.8	87.1	87.1	87.3	87.6	87.8
	docCOMET	82.3	82.5	78.0	82.6	79.8	**83.9**	75.6	80.4	74.2	81.1	79.7	82.2	80.7	82.2	82.2	82.5	82.9	83.2
	COMETQE	82.7	82.9	80.1	**83.6**	78.8	82.2	79.6	83.5	77.7	83.3	80.8	83.2	81.4	83.2	83.0	83.2	82.9	83.1
	docCOMETQE	80.1	80.5	75.5	**81.4**	74.1	79.7	73.2	81.3	68.9	81.1	78.5	80.8	78.7	80.8	80.6	81.0	80.7	81.1
	L.repetition	80.0	80.0	73.0	79.0	78.0	**83.0**	73.0	80.0	75.0	**83.0**	76.0	78.0	80.0	78.0	79.0	80.0	82.0	82.0
	Formality	65.0	67.0	46.0	62.0	58.0	**69.0**	56.0	66.0	53.0	68.0	19.0	64.0	38.0	64.0	64.0	65.0	66.0	67.0
	Pronouns	51.0	51.0	34.0	43.0	51.0	**61.0**	41.0	49.0	47.0	57.0	45.0	47.0	54.0	47.0	48.0	48.0	55.0	55.0
EN-DE	BLEU	31.3	32.7	20.9	31.7	21.9	**33.1**	15.6	29.0	14.1	29.2	28.1	29.7	29.0	31.8	27.3	28.3	29.4	30.4
	COMET	83.8	84.1	79.7	84.6	79.9	**85.1**	77.4	83.9	76.2	84.1	79.8	83.8	77.4	82.4	83.3	83.8	84.0	84.5
	docCOMET	79.1	79.4	74.3	79.9	75.3	**80.9**	70.5	78.0	70.2	79.1	74.3	78.8	71.9	77.4	78.4	78.9	79.4	79.9
	COMETQE	82.9	82.9	78.8	83.3	76.8	81.9	77.6	**83.4**	76.5	83.0	78.3	82.7	75.6	80.6	82.7	83.2	82.7	83.1
	docCOMETQE	80.8	81.0	74.3	**81.6**	72.3	80.1	68.9	81.2	67.9	80.8	76.5	80.7	73.1	78.9	80.7	81.4	80.9	81.5
	L. repetition	69.0	69.0	63.0	70.0	68.0	**76.0**	60.0	68.8	64.0	72.0	65.0	67.0	66.0	71.0	66.0	67.0	71.0	71.0
	Formality	65.0	67.0	62.0	70.0	67.0	**75.0**	56.0	68.0	58.0	70.0	10.0	60.0	09.0	30.0	56.0	57.0	70.0	71.0
	Pronouns	68.0	67.0	56.0	65.0	63.0	**73.0**	55.0	66.0	59.0	69.0	63.0	66.0	62.0	66.0	66.0	66.0	67.0	68.0
EN-FR	BLEU	41.0	**43.0**	27.2	40.3	31.1	42.9	21.2	37.2	20.5	38.7	36.8	38.3	37.7	39.1	37.2	38.3	39.2	40.1
	COMET	84.0	84.5	80.9	85.1	81.6	**85.7**	77.7	84.3	76.8	84.5	81.4	84.3	78.6	82.3	84.1	84.3	84.7	85.0
	docCOMET	80.0	80.6	76.4	81.2	77.8	**82.3**	71.6	79.3	71.4	80.1	76.9	80.1	73.9	78.0	80.1	80.1	80.9	81.3
	COMETQE	84.1	84.4	82.2	**85.3**	80.9	84.1	79.6	84.9	78.1	84.6	81.7	84.7	78.9	83.0	84.7	84.7	84.7	84.9
	docCOMETQE	81.7	82.5	78.1	**83.4**	77.1	82.2	71.1	82.5	68.9	82.2	80.1	82.4	76.6	80.6	82.6	82.4	82.8	83.2
	L. repetition	78.0	79.0	71.0	77.0	76.0	**81.0**	69.0	77.0	70.0	79.0	74.0	77.0	72.0	77.0	77.0	77.0	79.0	79.0
	Formality	75.0	74.0	66.0	75.0	71.0	**79.0**	60.0	75.0	61.0	76.0	18.0	71.0	11.0	51.0	66.0	66.0	77.0	77.0
	Pronouns	75.0	75.0	64.0	73.0	72.0	**79.0**	63.0	73.0	64.0	76.0	66.0	69.0	69.0	72.0	71.0	72.0	77.0	78.0
EN-RU	BLEU	24.2	24.9	16.1	24.4	11.7	**26.2**	14.7	23.8	15.6	25.4	21.5	22.4	23.0	24.2	21.1	21.8	22.7	23.7
	COMET	84.3	84.3	80.6	85.1	71.6	**85.8**	81.1	85.2	81.0	85.7	83.3	84.5	82.6	84.1	83.8	84.2	84.8	85.2
	docCOMET	79.7	79.8	75.5	80.5	66.9	**81.8**	75.8	80.6	76.0	81.2	78.2	79.6	77.8	79.5	79.1	79.5	80.4	80.8
	COMETQE	82.7	82.6	78.1	83.3	64.1	81.8	78.9	**83.5**	78.8	83.3	82.2	83.2	80.9	82.4	82.6	83.1	82.7	83.1
	docCOMETQE	80.0	79.8	71.7	81.0	65.6	79.4	72.2	81.2	72.7	**81.3**	79.6	80.7	78.3	80.1	79.9	80.7	80.3	81.1
	L. repetition	58.0	59.0	53.0	59.0	44.0	**64.0**	52.0	58.0	56.0	62.0	55.0	56.0	58.0	60.0	57.0	62.0	62.0	62.0
	Formality	56.0	56.0	47.0	57.0	39.0	**61.0**	48.0	58.0	48.0	60.0	31.0	50.0	21.0	42.0	51.0	60.0	59.0	60.0
EN-AR	BLEU	12.5	12.5	N/A	N/A	N/A	N/A	6.5	11.4	5.2	**13.4**	09.3	10.2	11.0	10.2	08.2	09.6	09.5	10.9
	COMET	81.3	81.2	N/A	N/A	N/A	N/A	77.2	82.1	75.0	**82.5**	80.1	81.5	78.0	81.6	77.9	80.9	79.3	81.8
	docCOMET	75.0	74.9	N/A	N/A	N/A	N/A	69.9	75.5	68.3	**76.2**	72.9	74.6	71.1	74.7	70.8	74.0	72.6	75.4
	COMETQE	79.1	78.7	N/A	N/A	N/A	N/A	74.0	**80.3**	70.8	79.5	78.9	**80.3**	75.8	**80.3**	75.9	79.3	76.3	79.3
	docCOMETQE	76.6	75.9	N/A	N/A	N/A	N/A	68.0	**78.4**	61.8	77.4	76.5	78.3	73.5	78.2	70.5	76.8	71.7	77.1
	L. repetition	55.0	55.0	N/A	N/A	N/A	N/A	48.0	54.0	53.0	**60.0**	52.0	53.0	54.0	53.0	50.0	51.0	57.0	58.0
	Pronouns	**51.0**	49.0	N/A	N/A	N/A	N/A	41.0	48.0	41.0	50.0	47.0	48.0	50.0	48.0	46.0	50.0	46.0	50.0
EN-KO	BLEU	20.6	20.9	14.2	20.8	9.7	20.3	12.1	20.9	13.6	**23.7**	16.9	18.7	20.3	22.4	16.4	18.3	19.4	21.4
	COMET	84.7	84.7	82.9	86.1	80.1	85.9	81.3	85.7	82.0	86.8	81.8	85.3	83.0	85.0	84.5	85.8	85.8	**86.9**
	docCOMET	79.1	79.0	76.4	80.4	74.4	80.9	74.8	80.1	76.0	81.6	75.2	79.1	77.1	79.5	78.2	79.7	80.4	**81.7**
	COMETQE	84.7	84.4	81.4	85.6	74.7	82.6	80.3	85.5	79.9	85.4	81.3	85.0	82.1	84.4	84.3	**85.7**	84.5	**85.7**
	docCOMETQE	80.9	80.4	75.8	82.3	67.7	79.2	72.7	82.1	73.4	82.2	78.6	81.5	79.2	81.3	80.0	82.3	80.7	**82.6**
	L. repetition	45.0	46.0	40.0	47.0	44.0	**52.0**	39.0	45.0	45.0	50.0	40.0	44.0	49.0	50.0	42.0	43.0	49.0	50.0
	Formality	26.0	24.0	23.0	27.0	26.0	**39.0**	20.0	28.0	27.0	38.0	04.0	32.0	13.0	26.0	15.0	13.0	33.0	35.0

Table 11: Detailed language-specific translation and discourse phenomena performance of all models (nllb=NLLB-3.3B, tower=TowerInstruct-13B, euro=EuroLLM-9B-Inst, gemma=Gemma3-12B-it, qwen=Qwen3-14B) using greedy (G) and quality-aware decoding (Q) on **TED2020 dataset** in both sentence-level (ctx= 0) and context-aware (ctx= 5) setups. N/A: not applicable as TowerInstruct-13B is not trained on Arabic. **Bold** highlights the best value per row. The random chance performance on discourse phenomena varies depending on the number of elements in the list of ambiguous words, which differs across languages.

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
<src_lang>: <src context 3> <tgt_lang>: <tgt context 3>
<src_lang>: <src context 4> <tgt_lang>: <tgt context 4>
<src_lang>: <src context 5> <tgt_lang>: <tgt context 5>
Given the provided parallel sentence pairs, translate the following <src_lang> sentence to
↪  <tgt_lang>.
<src_lang>: <src sentence> <tgt_lang>:
```

Figure 8: EuroLLM-9B-Inst prompt format.

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
<src_lang>: <src context 3> <tgt_lang>: <tgt context 3>
<src_lang>: <src context 4> <tgt_lang>: <tgt context 4>
<src_lang>: <src context 5> <tgt_lang>: <tgt context 5>
Given the provided parallel sentence pairs, translate the following <src_lang> sentence to
↪  <tgt_lang>. Do not give an explanation of your translation.
<src_lang>: <src sentence> <tgt_lang>:
```

Figure 9: Gemma3-12B-it and Qwen3-14B prompt format.

		nllb		tower				euro				gemma				qwen			
	Metric	ctx= 0		ctx= 0		ctx= 5		ctx= 0		ctx= 5		ctx= 0		ctx= 5		ctx= 0		ctx= 5	
		G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q	G	Q
EN-PT	BLEU	33.2	35.2	24.7	35.5	25.8	35.6	26.7	38.9	26.4	39.3	37.7	39.3	39.2	**40.5**	36.5	38.0	37.4	38.6
	COMET	78.8	79.5	79.3	83.3	79.2	83.2	78.8	82.8	78.1	83.2	82.4	83.8	83.1	84.0	83.5	83.8	83.9	**84.2**
	docCOMET	76.8	77.6	77.1	81.2	77.3	81.4	75.9	80.2	75.3	80.9	80.1	81.7	80.8	81.8	81.3	81.7	81.8	**82.2**
	COMETQE	75.7	76.5	75.0	**79.8**	73.7	78.4	74.8	79.7	73.5	79.1	77.9	79.6	78.1	79.2	79.2	79.5	79.2	79.5
	docCOMETQE	67.5	68.5	66.9	**73.2**	66.3	72.4	66.0	73.1	64.7	73.0	71.1	72.7	72.1	72.9	72.2	72.9	72.7	73.1
	L. repetition	77.0	76.0	75.0	79.0	78.0	83.0	79.0	82.0	79.0	84.0	79.0	80.0	84.0	84.0	80.0	83.0	85.0	**85.0**
	Formality	58.0	62.0	45.0	58.0	47.0	61.0	57.0	**66.0**	58.0	**66.0**	55.0	65.0	62.0	64.0	65.0	**65.0**	65.0	63.0
	Pronouns	49.0	50.0	20.0	35.0	42.0	49.0	45.0	45.0	46.0	**56.0**	47.0	48.0	46.0	45.0	49.0	44.0	44.0	47.0
EN-DE	BLEU	05.1	05.2	04.2	05.5	07.9	**12.6**	03.6	05.3	03.6	04.8	05.3	05.2	04.9	05.3	04.7	05.0	04.9	05.1
	COMET	48.1	47.5	48.2	50.4	58.2	**63.0**	48.3	50.8	48.2	50.8	50.3	50.8	47.9	49.7	49.9	50.4	50.4	50.6
	docCOMET	46.0	45.6	46.1	47.9	57.0	**61.6**	45.8	47.9	45.9	48.2	47.6	48.3	45.2	47.4	47.5	47.8	47.9	48.1
	COMETQE	77.1	76.9	74.5	**80.4**	58.6	65.2	74.0	80.2	67.4	75.9	77.4	80.1	70.9	77.1	79.6	**80.4**	79.6	80.2
	docCOMETQE	69.7	69.4	67.0	74.2	59.9	68.8	64.3	74.2	61.5	71.3	71.7	74.0	67.0	72.4	73.2	**74.5**	73.7	**74.5**
	L. repetition	26.0	27.0	33.0	29.0	22.0	23.0	32.0	29.0	30.0	29.0	30.0	30.0	**34.0**	33.0	28.0	28.0	29.0	28.0
	Formality	23.0	23.0	22.0	24.0	36.0	**37.0**	23.0	25.0	24.0	23.0	18.0	24.0	15.0	21.0	22.0	22.0	22.0	22.0
	Pronouns	26.0	23.0	19.0	28.0	22.0	14.0	22.0	26.0	26.0	23.0	27.0	27.0	25.0	26.0	28.0	27.0	26.0	**29.0**
EN-FR	BLEU	23.3	32.8	24.5	35.3	25.7	36.9	24.2	36.5	23.3	32.8	36.0	38.0	36.2	**38.8**	33.8	35.0	34.9	36.3
	COMET	74.0	79.7	77.0	81.2	76.6	81.3	74.6	80.2	74.0	79.7	79.4	81.2	78.3	81.0	80.4	80.8	81.2	**81.7**
	docCOMET	72.7	78.0	75.3	79.6	75.6	80.1	72.5	78.3	72.7	78.0	77.9	79.8	77.0	79.6	78.9	79.2	79.7	**80.3**
	COMETQE	74.0	80.1	77.1	**81.8**	75.5	81.1	75.5	81.5	74.0	80.1	79.6	81.3	78.0	80.7	81.3	81.8	81.2	81.7
	docCOMETQE	65.1	73.5	68.2	74.7	68.3	74.6	64.8	73.9	65.1	73.5	72.8	74.1	72.4	74.0	74.0	74.5	74.3	**75.0**
	L. repetition	66.0	72.0	69.0	75.0	72.0	**80.0**	65.0	73.0	66.0	72.0	73.0	75.0	73.0	77.0	73.0	74.0	79.0	78.0
	Formality	49.0	58.0	53.0	57.0	57.0	61.0	50.0	**61.0**	49.0	58.0	56.0	64.0	57.0	**68.0**	61.0	62.0	63.0	65.0
	Pronouns	47.0	53.0	47.0	47.0	50.0	**60.0**	44.0	50.0	47.0	53.0	51.0	53.0	57.0	**60.0**	50.0	49.0	58.0	**60.0**
	Verb form	37.0	**49.0**	27.0	43.0	30.0	45.0	35.0	**49.0**	37.0	**49.0**	47.0	48.0	44.0	47.0	45.0	46.0	45.0	45.0
EN-RU	BLEU	20.6	20.7	14.8	22.5	13.7	23.2	14.7	22.3	15.1	23.5	22.3	23.5	23.8	**24.6**	20.0	21.6	22.1	23.2
	COMET	76.1	76.2	76.2	81.2	73.8	81.5	76.4	81.1	76.0	81.6	81.1	82.0	81.0	**82.1**	80.2	80.7	81.6	81.9
	docCOMET	74.0	74.1	73.7	78.6	72.5	79.6	73.9	78.6	74.0	79.4	78.4	79.6	78.7	**79.8**	77.6	78.2	79.3	79.7
	COMETQE	75.8	75.5	71.7	79.0	67.5	77.5	72.6	79.3	72.2	78.8	78.7	79.4	77.5	78.7	78.6	**79.5**	78.7	79.2
	docCOMETQE	67.0	67.0	61.7	71.6	59.5	70.6	62.5	71.7	63.6	71.5	71.6	**72.2**	70.6	71.8	70.7	72.1	71.3	**72.2**
	L. repetition	63.0	53.0	50.0	63.0	65.0	72.0	42.0	60.0	64.0	**75.0**	46.0	49.0	72.0	73.0	48.0	48.0	72.0	**75.0**
	Formality	47.0	48.0	39.0	48.0	34.0	**55.0**	43.0	53.0	49.0	52.0	48.0	51.0	51.0	53.0	41.0	44.0	48.0	30.0
	Verb form	32.0	34.0	26.0	36.0	21.0	**38.0**	22.0	37.0	28.0	37.0	35.0	**38.0**	33.0	**38.0**	28.0	33.0	32.0	33.0
EN-AR	BLEU	17.5	16.9	N/A	N/A	N/A	N/A	00.3	00.5	10.3	**20.6**	00.4	00.5	00.4	00.5	00.4	00.4	00.5	00.5
	COMET	77.8	77.1	N/A	N/A	N/A	N/A	50.1	52.4	75.3	**81.7**	53.4	53.7	50.6	52.3	50.7	52.5	50.7	52.4
	docCOMET	75.2	74.3	N/A	N/A	N/A	N/A	46.6	48.7	72.7	**79.2**	49.5	50.0	47.2	49.0	47.2	48.9	47.4	49.0
	COMETQE	72.7	71.0	N/A	N/A	N/A	N/A	65.7	74.9	66.8	75.1	74.6	**76.1**	69.1	72.9	70.1	74.8	69.7	74.4
	docCOMETQE	64.7	63.1	N/A	N/A	N/A	N/A	55.3	68.1	57.5	67.9	68.0	**69.4**	65.4	68.0	61.6	67.3	61.6	67.2
	L. repetition	63.0	60.0	N/A	N/A	N/A	N/A	33.0	60.0	58.0	**75.0**	50.0	54.0	51.0	59.0	53.0	60.0	62.0	69.0
	Pronouns	55.0	43.0	N/A	N/A	N/A	N/A	36.0	44.0	36.0	54.0	39.0	**56.0**	41.0	41.0	49.0	53.0	51.0	**56.0**
EN-KO	BLEU	22.2	22.1	16.2	25.2	12.0	20.1	15.0	25.8	17.0	28.6	25.3	27.6	27.8	**30.6**	21.6	24.9	23.5	26.7
	COMET	80.1	79.6	80.8	85.0	77.4	83.4	79.2	84.6	79.9	85.7	83.8	**86.5**	84.9	86.3	83.9	85.5	84.8	86.2
	docCOMET	77.2	76.7	77.2	81.7	75.1	80.9	75.8	81.5	76.9	82.8	80.1	82.0	81.7	**83.3**	80.3	81.9	81.6	83.0
	COMETQE	79.4	78.7	75.9	81.9	69.5	77.9	75.1	81.9	75.4	82.1	79.9	81.9	80.4	82.3	80.4	**82.7**	80.6	82.5
	docCOMETQE	70.2	70.3	67.7	75.2	63.7	73.6	65.3	74.6	66.6	75.2	73.4	75.2	74.7	75.8	72.8	75.8	73.2	**76.0**
	L. repetition	44.0	32.0	22.0	37.0	34.0	43.0	23.0	41.0	38.0	57.0	39.0	44.0	55.0	**62.0**	32.0	35.0	48.0	50.0
	Formality	27.0	27.0	10.0	30.0	25.0	32.0	09.0	22.0	09.0	34.0	44.0	38.0	33.0	**48.0**	17.0	07.0	40.0	31.0

Table 12: Detailed language-specific translation and discourse phenomena performance of all models (nllb=NLLB-3.3B, tower=TowerInstruct-13B, euro=EuroLLM-9B-Inst, gemma=Gemma3-12B-it, qwen=Qwen3-14B) using greedy (G) and quality-aware decoding (Q) on **WMT24++ dataset** in both sentence-level (ctx= 0) and context-aware (ctx= 5) setups. N/A: not applicable as TowerInstruct-13B is not trained on Arabic. **Bold** highlights the best value per row.

	BLEU	COMET	docCOMET	COMETQE	docCOMETQE	L.repetition	Formality	Pronouns
nllb Greedy	55.2	87.1	82.2	81.5	75.4	87.0	75.0	45.0
nllb Beam	55.9	87.2	82.1	81.8	76.4	85.0	75.0	48.0

Table 13: Translation and discourse phenomena performance of NLLB-3.3B model using greedy decoding and beam search decoding (beam=5).

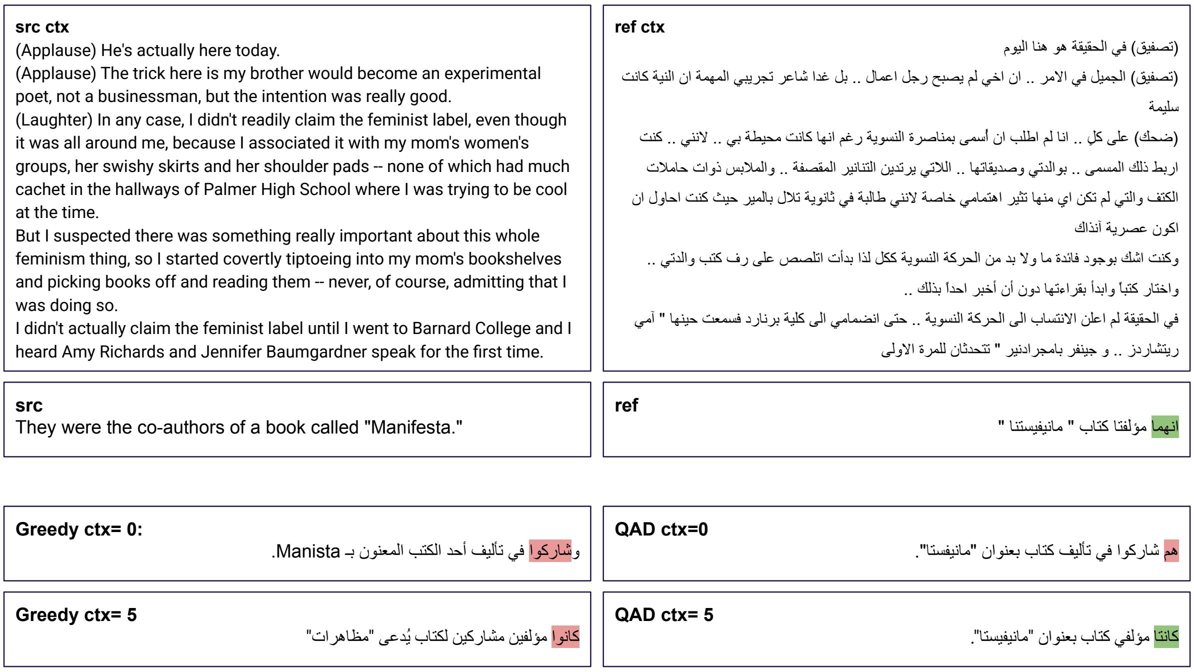
We present the participants with 25 samples including the following data:

- The **source context** which was given to the translation model, which are (up to 5) previous sentences in the source document.
- The English source sentence.
- The **output context** which was given to the translation model, which are (up to 5) previous sentences in the output document.
- **output 1**: the output of the first system
- **output 2**: the output of the second system

Annotators are asked to assess the following:

- **Semantic difference**: Rate the semantic difference of the two outputs on a scale of 1 to 5, ignoring differences in wording. Consider whether they convey the same meaning.
  - 1: the two sentences convey the same meaning.
  - 5: the two sentences convey completely different meanings.
- **Pronoun resolution**: Does the source sentence contain an ambiguous pronoun (a pronoun whose referent is unclear or not explicitly mentioned), and what is it?
  - If yes, is it correctly translated in output 1?
  - If yes, is it correctly translated in output 2?
- **Lexical repetition**: Does the source sentence contain an entity (e.g., noun, occupation) previously mentioned in the source context, and what is the entity?
  - If yes, is it translated consistently with its previous translation in the output context in output 1?
  - If yes, is it translated consistently with its previous translation in the output context in output 2?
- **Formality**: Does the source sentence exhibit a formality phenomenon (e.g., addressing someone formally or expressing respect), and what is the word that exhibits the phenomenon?
  - If yes, is it handled in the output 1?
  - If yes, is it handled in the output 2?
- **Verb form**: Does the source sentence contain an ambiguous verb that can have different forms depending on the gender or formality level of the subject, and what is the verb?
  - If yes, is it correctly translated in output 1?
  - If yes, is it correctly translated in output 2?
- **General comment (optional)**: Provide comments or observations about the two outputs. Highlight strengths, weaknesses, or notable phenomena (e.g., mistranslation, cultural adaptation, or syntactic errors). Please also highlight other linguistic phenomena we may have missed in the categories provided.
- **Preference**: Which output do you prefer? (output 1, output 2, equally good, equally bad)

Table 14: Human assessment details.



src ctx
(Applause) He's actually here today.
(Applause) The trick here is my brother would become an experimental poet, not a businessman, but the intention was really good.
(Laughter) In any case, I didn't readily claim the feminist label, even though it was all around me, because I associated it with my mom's women's groups, her swishy skirts and her shoulder pads -- none of which had much cachet in the hallways of Palmer High School where I was trying to be cool at the time.
But I suspected there was something really important about this whole feminism thing, so I started covertly tiptoeing into my mom's bookshelves and picking books off and reading them -- never, of course, admitting that I was doing so.
I didn't actually claim the feminist label until I went to Barnard College and I heard Amy Richards and Jennifer Baumgardner speak for the first time.

ref ctx
(تصفيق) في الحقيقة هو هنا اليوم
(تصفيق) الجميل في الامر .. ان اخي لم يصبح رجل اعمال .. بل غدا شاعر تجريبي المهمة ان النية كانت سليمة
(ضحك) على كلٍ .. انا لم اطلب ان أسمى بمناصرة النسوية رغم انها كانت محيطة بي .. لانني .. كنت اربط ذلك المسمى .. بوالدتي وصديقاتها .. اللاتي يرتدين التنانير المقصفة .. والملابس ذوات حاملات الكتف والتي لم تكن اي منها تثير اهتمامي خاصة لانني طالبة في ثانوية تلال بالمير حيث كنت احاول ان اكون عصرية آنذاك
وكنت اشك بوجود فائدة ما ولا بد من الحركة النسوية ككل لذا بدأت اتلصص على رف كتب والدتي .. واختار كتباً وابدأ بقراءتها دون أن أخبر احداً بذلك ..
في الحقيقة لم اعلن الانتساب الى الحركة النسوية .. حتى انضمامي الى كلية برنارد فسمعت حينها " آمي ريتشاردز .. و جينفر بامجرادنير " تتحدثان للمرة الاولى

src
They were the co-authors of a book called "Manifesta."

ref
انهما مؤلفتا كتاب " مانيفيستنا "

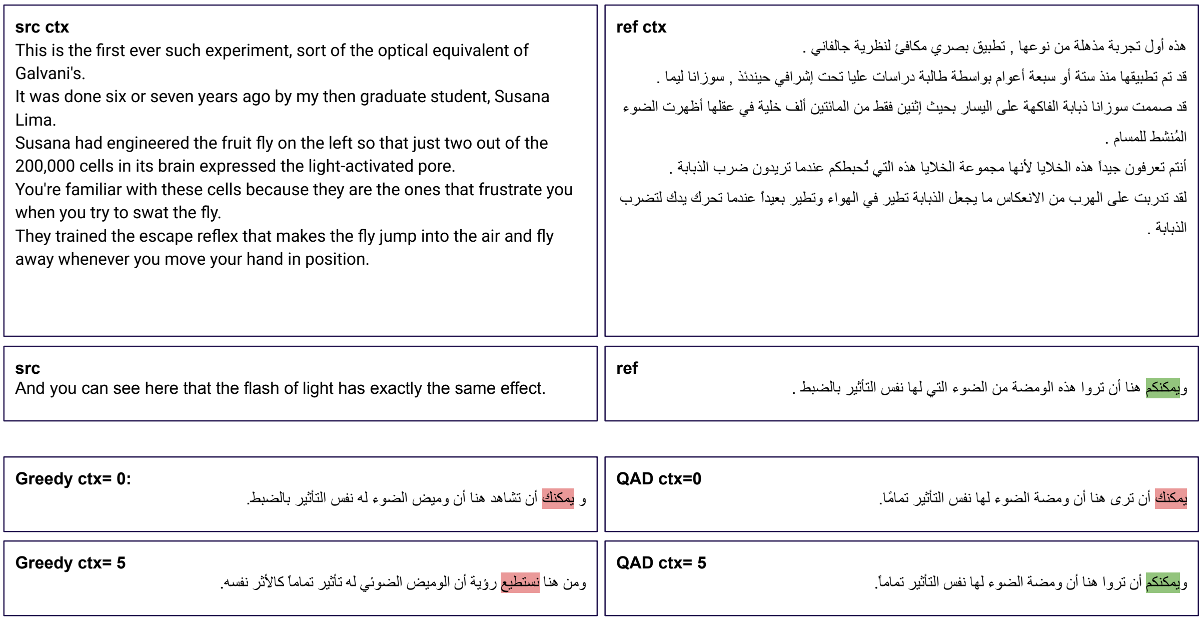
Greedy ctx= 0:
وشاركوا في تأليف أحد الكتب المعنون بـ Manista.

QAD ctx=0
هم شاركوا في تأليف كتاب بعنوان "مانيفستا".

Greedy ctx= 5
كانوا مؤلفين مشاركين لكتاب يُدعى "مظاهرات"

QAD ctx= 5
كانتا مؤلفي كتاب بعنوان "مانيفيستا".

Figure 10: Dual pronouns in Arabic



src ctx
This is the first ever such experiment, sort of the optical equivalent of Galvani's.
It was done six or seven years ago by my then graduate student, Susana Lima.
Susana had engineered the fruit fly on the left so that just two out of the 200,000 cells in its brain expressed the light-activated pore.
You're familiar with these cells because they are the ones that frustrate you when you try to swat the fly.
They trained the escape reflex that makes the fly jump into the air and fly away whenever you move your hand in position.

ref ctx
هذه أول تجربة مذهلة من نوعها , تطبيق بصري مكافئ لنظرية جالفاني .
قد تم تطبيقها منذ ستة أو سبعة أعوام بواسطة طالبة دراسات عليا تحت إشرافي حينذئذ , سوزانا ليما .
قد صممت سوزانا ذبابة الفاكهة على اليسار بحيث إثنين فقط من المائتين ألف خلية في عقلها أظهرت الضوء المُنشط للمسام .
أنتم تعرفون جيداً هذه الخلايا لأنها مجموعة الخلايا هذه التي تُحبطكم عندما تريدون ضرب الذبابة .
لقد تدربت على الهرب من الانعكاس ما يجعل الذبابة تطير في الهواء وتطير بعيداً عندما تحرك يدك لتضرب الذبابة .

src
And you can see here that the flash of light has exactly the same effect.

ref
ويمكنكم هنا أن تروا هذه الومضة من الضوء التي لها نفس التأثير بالضبط .

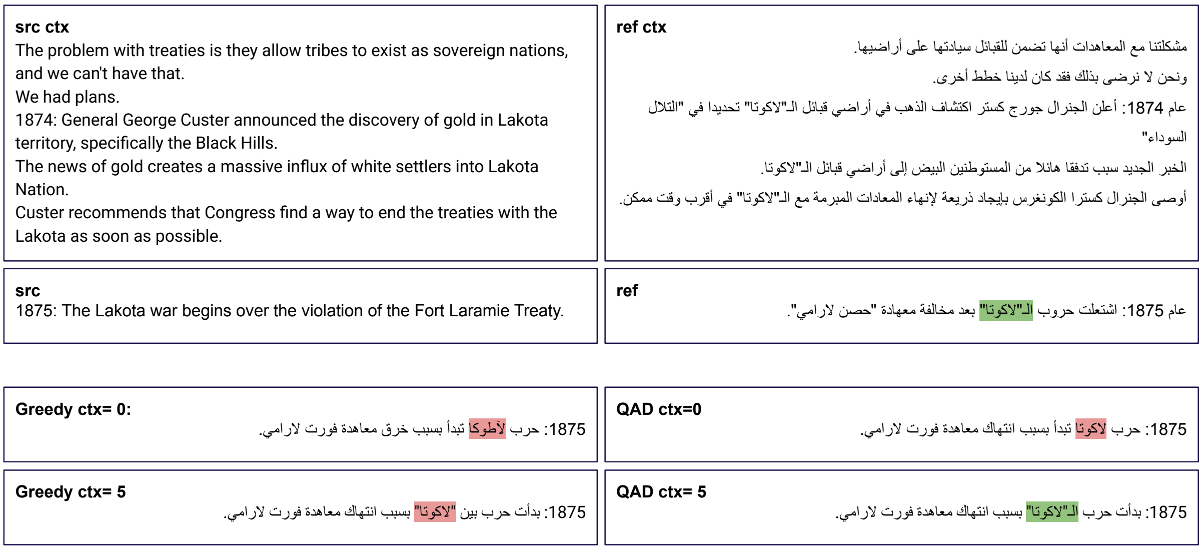
Greedy ctx= 0:
و يمكنك أن تشاهد هنا أن وميض الضوء له نفس التأثير بالضبط.

QAD ctx=0
يمكنك أن ترى هنا أن ومضة الضوء لها نفس التأثير تمامًا.

Greedy ctx= 5
ومن هنا نستطيع رؤية أن الوميض الضوئي له تأثير تماماً كالأثر نفسه.

QAD ctx= 5
ويمكنكم أن تروا هنا أن ومضة الضوء لها نفس التأثير تماماً.

Figure 11: Plural pronouns in Arabic.



src ctx
The problem with treaties is they allow tribes to exist as sovereign nations, and we can't have that.
We had plans.
1874: General George Custer announced the discovery of gold in Lakota territory, specifically the Black Hills.
The news of gold creates a massive influx of white settlers into Lakota Nation.
Custer recommends that Congress find a way to end the treaties with the Lakota as soon as possible.

ref ctx
مشكلتنا مع المعاهدات أنها تضمن للقبائل سيادتها على أراضيها.
ونحن لا نرضى بذلك فقد كان لدينا خطط أخرى.
عام 1874: أعلن الجنرال جورج كستر اكتشاف الذهب في أراضي قبائل الـ"لاكوتا" تحديدا في "التلال السوداء"
الخبر الجديد سبب تدفقا هائلا من المستوطنين البيض إلى أراضي قبائل الـ"لاكوتا.
أوصى الجنرال كسترا الكونغرس بإيجاد ذريعة لإنهاء المعادات المبرمة مع الـ"لاكوتا" في أقرب وقت ممكن.

src
1875: The Lakota war begins over the violation of the Fort Laramie Treaty.

ref
عام 1875: اشتعلت حروب الـ"لاكوتا" بعد مخالفة معهادة "حصن لارامي".

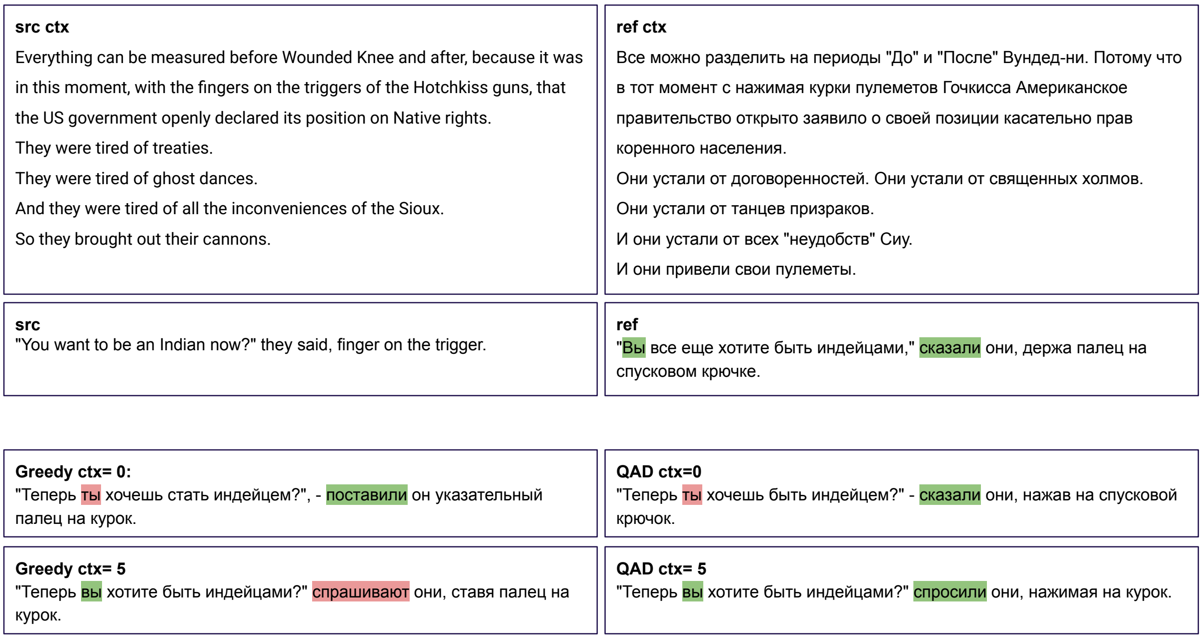
Greedy ctx= 0:
1875: حرب لآطوكا تبدأ بسبب خرق معاهدة فورت لارامي.

QAD ctx=0
1875: حرب لاكوتا تبدأ بسبب انتهاك معاهدة فورت لارامي.

Greedy ctx= 5
1875: بدأت حرب بين "لاكوتا" بسبب انتهاك معاهدة فورت لارامي.

QAD ctx= 5
1875: بدأت حرب الـ"لاكوتا" بسبب انتهاك معاهدة فورت لارامي.

Figure 12: Lexical repetition in Arabic.



src ctx
Everything can be measured before Wounded Knee and after, because it was in this moment, with the fingers on the triggers of the Hotchkiss guns, that the US government openly declared its position on Native rights.
They were tired of treaties.
They were tired of ghost dances.
And they were tired of all the inconveniences of the Sioux.
So they brought out their cannons.

ref ctx
Все можно разделить на периоды "До" и "После" Вундед-ни. Потому что в тот момент с нажимая курки пулеметов Гочкисса Американское правительство открыто заявило о своей позиции касательно прав коренного населения.
Они устали от договоренностей. Они устали от священных холмов.
Они устали от танцев призраков.
И они устали от всех "неудобств" Сиу.
И они привели свои пулеметы.

src
"You want to be an Indian now?" they said, finger on the trigger.

ref
"Вы все еще хотите быть индейцами," сказали они, держа палец на спусковом крючке.

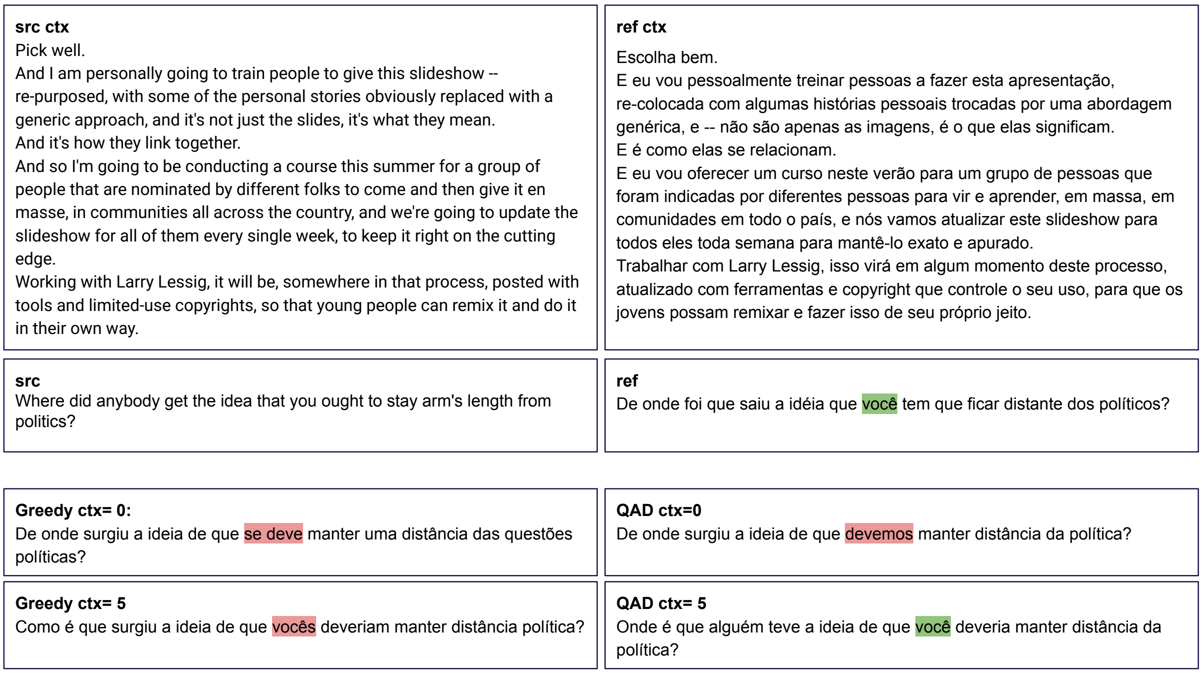
Greedy ctx= 0:
"Теперь ты хочешь стать индейцем?", - поставили он указательный палец на курок.

QAD ctx=0
"Теперь ты хочешь быть индейцем?" - сказали они, нажав на спусковой крючок.

Greedy ctx= 5
"Теперь вы хотите быть индейцами?" спрашивают они, ставя палец на курок.

QAD ctx= 5
"Теперь вы хотите быть индейцами?" спросили они, нажимая на курок.

Figure 13: Plural pronoun and verb form in Russian.



src ctx
Pick well.
And I am personally going to train people to give this slideshow -- re-purposed, with some of the personal stories obviously replaced with a generic approach, and it's not just the slides, it's what they mean.
And it's how they link together.
And so I'm going to be conducting a course this summer for a group of people that are nominated by different folks to come and then give it en masse, in communities all across the country, and we're going to update the slideshow for all of them every single week, to keep it right on the cutting edge.
Working with Larry Lessig, it will be, somewhere in that process, posted with tools and limited-use copyrights, so that young people can remix it and do it in their own way.

ref ctx
Escolha bem.
E eu vou pessoalmente treinar pessoas a fazer esta apresentação, re-colocada com algumas histórias pessoais trocadas por uma abordagem genérica, e -- não são apenas as imagens, é o que elas significam.
E é como elas se relacionam.
E eu vou oferecer um curso neste verão para um grupo de pessoas que foram indicadas por diferentes pessoas para vir e aprender, em massa, em comunidades em todo o país, e nós vamos atualizar este slideshow para todos eles toda semana para mantê-lo exato e apurado.
Trabalhar com Larry Lessig, isso virá em algum momento deste processo, atualizado com ferramentas e copyright que controle o seu uso, para que os jovens possam remixar e fazer isso de seu próprio jeito.

src
Where did anybody get the idea that you ought to stay arm's length from politics?

ref
De onde foi que saiu a idéia que você tem que ficar distante dos políticos?

Greedy ctx= 0:
De onde surgiu a ideia de que se deve manter uma distância das questões políticas?

QAD ctx=0
De onde surgiu a ideia de que devemos manter distância da política?

Greedy ctx= 5
Como é que surgiu a ideia de que vocês deveriam manter distância política?

QAD ctx= 5
Onde é que alguém teve a ideia de que você deveria manter distância da política?

Figure 14: Formality and pronoun in Brazilian-Portuguese.