Unlocking Latent Discourse Translation in LLMs Through Quality-Aware Decoding

Anonymous ACL submission

Abstract

Large language models (LLMs) have emerged as strong contenders in machine translation. Yet, they often fall behind specialized neural machine translation systems in addressing 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 for document-level translation. We demonstrate that discourse knowledge is encoded within LLMs and propose the use of quality-aware decoding (QAD) 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 not only for sentence-level but also 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). However, document-level translation introduces extra challenges, 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, hence inspiring several works that try to trace and understand such patterns and differences in neural machine translation (NMT). 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 LLMs' 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 the performance of LLMs in document-level translation with respect to different discourse phenomena. Inspired by Fernandes et al. (2023), we measure models' performance on four phenomena: lexical cohesion, 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 fully exploited by greedy decoding. We thus experiment with quality-aware decoding (Fernandes et al., 2022) and find that it indeed helps improve the discourse phenomena performance of LLMs. We validate our findings through extensive

Lexical cohesion	**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.

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).

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 terms of 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 cohesion, 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 as proposed by (Fernandes et al., 2023):

Lexical cohesion. Entities that are mentioned multiple times across the 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. Linguistic indicators such as pronouns and honorifics are used when addressing someone formally or expressing respect.

Verb form. Verbs should be translated according to the tense, gender, tone, mood, and cohesion of the document.

Examples of the phenomena highlighting the ambiguous words and their supporting context are presented in Table 1.

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 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 work we focus on MBR decoding.

A machine translation model defines a probability distribution $p(y|x, \theta)$ over a set of hypothesis $\mathcal{Y}$.

[1] All code and data will be released upon acceptance.

	Lexical cohesion	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 2: 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 sentence count can be less than the sum of phenomena counts because we can have multiple phenomena per sentence.

MAP decoding, such as greedy decoding, aims to maximize the probability of generated hypothesis:

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

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 $\mathcal{H}$. It selects:

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

We experiment with different choices for the utility function, including lexical, pretrained, and discourse-specific metrics for translation evaluation. We discuss these in more detail in §5.1.

3 Experiments

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 documents 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 20K subset of TED2020 data (Reimers and Gurevych, 2020) available in OPUS (Tiedemann, 2012). We also experiment on WMT24++ dataset (Deutsch et al., 2025) (results are in appendix A). 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 Table 2.

3.2 Models

We experiment on strong LLMs for translation, including TowerInstruct-13B (Alves et al., 2024); an instruction-tuned translation-specialized LLM based on Llama2-13B (Touvron et al., 2023), and EuroLLM-9B-Inst (Martins et al., 2024); a multilingual LLM trained from scratch on all European Union languages and additional relevant ones. We also experiment on NLLB-3.3B (Costa-jussà et al., 2022) as an encoder-decoder baseline.

3.3 Inference

We experiment with two decoding setups: **greedy** decoding, 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). We use **BLEU** score (Papineni et al., 2002) as the utility function for all our experiments unless stated otherwise. We conducted preliminary experiments on different prompting formats for each model and present only the best setup in this work. For TowerInstruct-13B and EuroLLM-9B-Inst, we employ context-aware

	NLLB-3.3B		TowerInstruct-13B		EuroLLM-9B-Inst	
	Greedy	QAD	Greedy	QAD	Greedy	QAD
BLEU	55.2	**58.2**	41.0	57.4	25.9	52.1
COMET	87.1	87.4	86.0	**89.6**	80.4	87.8
COMETQE	81.5	81.6	79.1	**82.0**	76.0	81.9
Lexical cohesion	87.0	85.0	85.0	**90.0**	79.0	89.0
Formality	75.0	**76.0**	66.0	**76.0**	56.0	75.0
Pronouns	45.0	47.0	50.0	**60.0**	40.0	48.0

Table 3: Translation and discourse phenomena performance of the three models using greedy and QAD setups on DELA dataset. **Bold** highlights the best value per row. The numbers presented for phenomena are F1 accuracies (details in §3.4.2). The results demonstrate that QAD enhances the performance of LLMs.

prompting with the context being (up to) 5 previous source-target pairs in the same document (prompt formats in Appendix C). For NLLB-3.3B, since the model has only been trained on sentence-level data, we conduct inference at the sentence level.

3.4 Evaluation

We measure both the overall translation performance and the discourse phenomena performance. We also include an LLM-based evaluation for completeness.

3.4.1 Overall Translation Evaluation

We use a lexical metric, BLEU (Papineni et al., 2002), a reference-based pretrained metric, COMET[2] (Rei et al., 2022a) and its reference-free variant, COMETQE[3] (Rei et al., 2022b).

3.4.2 Discourse Phenomena Evaluation

We measure the F1 accuracy of tagged words with discourse phenomena in the reference, existing and also being tagged in the hypothesis. To do so, we utilize the multilingual discourse-aware benchmark (MuDA) for discourse phenomena evaluation (Fernandes et al., 2023). The tagging of words is done automatically using a predefined language-specific list of pronouns, verb forms, and formality indicators. For lexical cohesion, the tagging is done by obtaining source-target word alignments, if an alignment pair occurs more than a specific number of times (three in our experiments, following MuDA), the word is tagged for lexical cohesion.

3.4.3 LLM-Based Evaluation

Evaluating LLMs automatically has become increasingly difficult due to their rapid advancements. Consequently, the use of language models for the automatic assessment of 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 then assigns a rating between 1 and 5, accompanied by an explanation of the decision. Since we use different prompting setups for our models (§3.3), the instructions provided for 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. Sustainability statement of all experiments in this paper is in Appendix E.

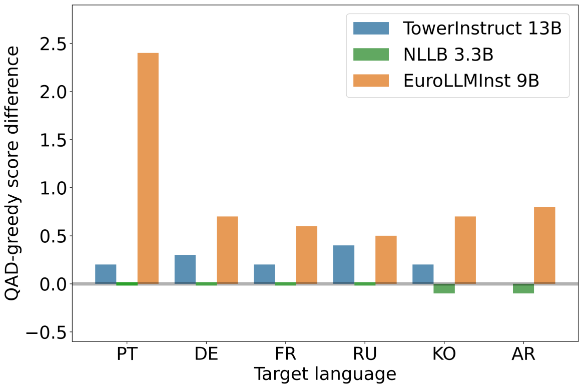

Figure 1: Difference between QAD and greedy LLM-as-a-judge scores on TED2020 data. The plot demonstrates that QAD improves the performance of LLMs.

4 Results

4.1 DELA-Data Results

In Table 3, we present the results on the DELA corpus. We see that TowerInstruct-13B and EuroLLM-9B-Inst fall behind NLLB-3.3B in translation and discourse phenomena performance when using

[2] https://huggingface.co/Unbabel/wmt22-comet-da
[3] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Language Pair	Metric	**NLLB-3.3B**		**TowerInstruct-13B**		**EuroLLM-9B-Inst**	
		Greedy	QAD	Greedy	QAD	Greedy	QAD
EN-PT	BLEU	40.4	41.8	30.4	**42.5**	21.0	38.9
	COMET	87.0	87.2	84.6	**88.2**	81.1	87.2
	COMETQE	82.7	82.9	78.8	82.2	77.7	**83.3**
	Lexical Cohesion	80.0	80.0	78.0	**83.0**	75.0	**83.0**
	Formality	65.0	67.0	58.0	**69.0**	53.0	68.0
	Pronouns	51.0	51.0	51.0	**61.0**	47.0	57.0
EN-DE	BLEU	31.3	32.7	21.9	**33.1**	14.1	29.2
	COMET	83.8	84.1	79.9	**85.1**	76.2	84.1
	COMETQE	82.9	82.9	76.8	81.9	76.5	**83.0**
	Lexical Cohesion	69.0	69.0	68.0	**76.0**	64.0	72.0
	Formality	65.0	67.0	67.0	**75.0**	58.0	70.0
	Pronouns	68.0	67.0	63.0	**73.0**	59.0	69.0
EN-FR	BLEU	41.0	**43.0**	31.1	42.9	20.5	38.7
	COMET	84.0	84.5	81.6	**85.7**	76.8	84.5
	COMETQE	84.1	84.4	80.9	84.1	78.1	**84.6**
	Lexical Cohesion	78.0	79.0	76.0	**81.0**	70.0	79.0
	Formality	75.0	74.0	71.0	**79.0**	61.0	76.0
	Pronouns	75.0	75.0	72.0	**79.0**	64.0	76.0
EN-RU	BLEU	24.2	24.9	11.7	**26.2**	15.6	25.4
	COMET	84.3	84.3	71.6	**85.8**	81.0	85.7
	COMETQE	82.7	82.6	64.1	81.8	78.8	**83.3**
	Lexical Cohesion	58.0	59.0	44.0	**64.0**	56.0	62.0
	Formality	56.0	56.0	39.0	**61.0**	48.0	60.0
EN-AR	BLEU	12.5	12.5	N/A	N/A	5.2	**13.4**
	COMET	81.3	81.2	N/A	N/A	75.0	**82.5**
	COMETQE	79.1	78.7	N/A	N/A	70.8	**79.5**
	Lexical Cohesion	55.0	55.0	N/A	N/A	53.0	**60.0**
	Pronouns	**51.0**	49.0	N/A	N/A	41.0	50.0
EN-KO	BLEU	20.6	20.9	9.7	20.3	13.6	**23.7**
	COMET	84.7	84.7	80.1	85.9	82.0	**86.8**
	COMETQE	84.7	84.4	74.7	82.6	79.9	**85.4**
	Lexical Cohesion	45.0	46.0	44.0	**52.0**	45.0	50.0
	Formality	26.0	24.0	26.0	**39.0**	27.0	38.0

Table 4: Translation and discourse phenomena performance of the three models using greedy and QAD setups on TED2020 dataset. N/A: not applicable as TowerInstruct-13B is not trained on Arabic. **Bold** highlights the best value per row. The results demonstrate that QAD enhances the performance of LLMs. The random chance performance varies depending on the number of elements in the list of ambiguous words, which differs across languages.

	BLEU	ChrF	COMET	COMETQE	Lexical cohesion	Formality	Pronouns
TowerInstruct-13B							
Greedy	41.0	55.8	86.0	79.1	85.0	66.0	50.0
QAD (BLEU)	**57.4**	76.3	89.6	82.0	**90.0**	76.0	60.0
QAD (ChrF)	55.8	**76.9**	89.7	82.2	**90.0**	**77.0**	61.0
QAD (COMET)	54.2	75	**90.9**	83.1	89.0	76.0	**62.0**
QAD (LC)	41.3	67.5	84.6	79.6	85.0	64.0	49.0
QAD (DiscoScore)	55.3	75.1	89.4	81.8	89.0	76.0	57.0
Fusion (COMETQE)	41.6	67.8	89.1	**85.7**	86.0	67.0	46.0
APE	44.3	68.4	87.7	82.0	84.0	69.0	47.0

Table 5: Translation and discourse phenomena performance of different decoding setups using TowerInstruct-13B on DELA data. **Bold** highlights the best value per column.

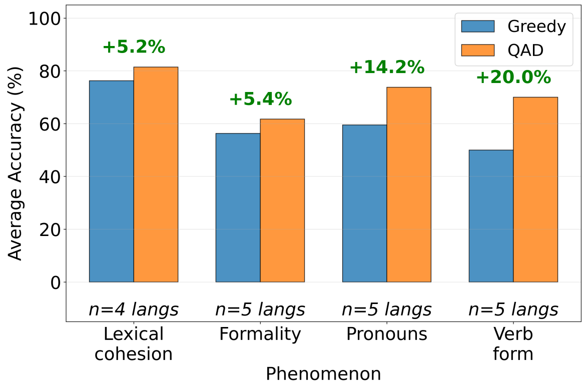

Figure 2: Human-annotated accuracy of greedy and QAD outputs in handling discourse phenomena, averaged across all languages where the phenomena occur (number of languages is shown at the bottom of the plot). Arabic is excluded from this plot to avoid model-specific biases.

greedy decoding. Interestingly, we observe that using QAD significantly improves both overall translation and discourse phenomena handling of LLMs allowing them to outperform NLLB-3.3B.

4.2 TED2020 and WMT24++ Results

In Table 4 we show the results on the TED2020 dataset for all language pairs. WMT24++ results are deferred to Appendix A as they evidence similar overall trends. The results highlight the substantial improvements in discourse and translation performance of LLMs using QAD across all language pairs. Moreover, TowerInstruct-13B outperforms other tested models, highlighting the effectiveness of translation finetuning in encoding discourse knowledge in LLMs.

Additionally, we present the differences between greedy and QAD scores from the LLM-as-a-judge evaluation for TED2020 data in Figure 1, WMT24++ results are in Appendix A. The results show that QAD enhances the performance of both LLMs (TowerInstruct-13B, EuroLLM-9B-Inst) but not NLLB-3.3B, which is consistent with the findings from automatic metrics.

5 Analysis

5.1 Inference Setup Ablation

We perform an ablation study on 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. For those metrics, we perform QAD using MBR with 50 samples 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 perform QAD using MBR with 20 samples generated using nucleus sampling (p=0.9).[4]

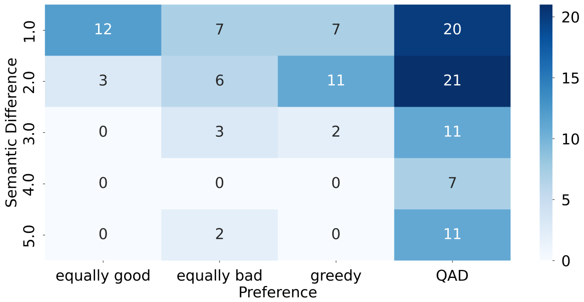

Figure 3: Semantic difference vs. preference summed over all languages (except Arabic).

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 leveraging XTOWER (Treviso et al., 2024) and XCOMET (Guerreiro et al., 2024), as used by the IT-Unbabel team in their submission to the quality estimation shared task at WMT24 (Zerva et al., 2024).

We assess the methods based on their translation and discourse phenomena performance, as shown in Table 5. Our analysis reveals that QAD outperforms other inference approaches, including fusion and APE. Notably, translation metrics serve as more effective utility functions compared to discourse-specific metrics. Among translation metrics, lexical measures (BLEU, ChrF) slightly outperform the pretrained COMET, though overall performance remains comparable. To further understand the lexical changes of the different inference setups, we analyze the distribution of edit

[4] We used 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.

	Lexical Cohesion	Formality	Pronouns	Verb form	Total	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 sentence count can be less than the sum of phenomena counts because we can have multiple phenomena per sentence.

operations (insert, delete, substitute, shift) in their outputs compared to greedy outputs to understand the lexical choices needed to improve the discourse and translation performance. The analysis in Figure 4 focuses on sentences tagged with discourse phenomena using MuDA (Fernandes et al., 2023). We show the percentage and absolute counts of edit operations for each setup compared to the greedy outputs, along with the overall edit rate on top of the bar plots. The analysis highlights that substitutions are the most frequent edit operation, followed by deletions, insertions, and shifts. Additionally, the findings indicate that an optimal level of edit operations produces strong results, as demonstrated by the utility functions BLEU, ChrF, COMET, and DiscoScore. However, deviations from this balance, whether through fewer edits (LC) or excessive edits (fusion), lead to poorer performance. These observations align with the performance scores reported in Table 5. Overall, this analysis highlights that among the experimented setups, **QAD with translation metrics is the best setup to improve discourse performance**.

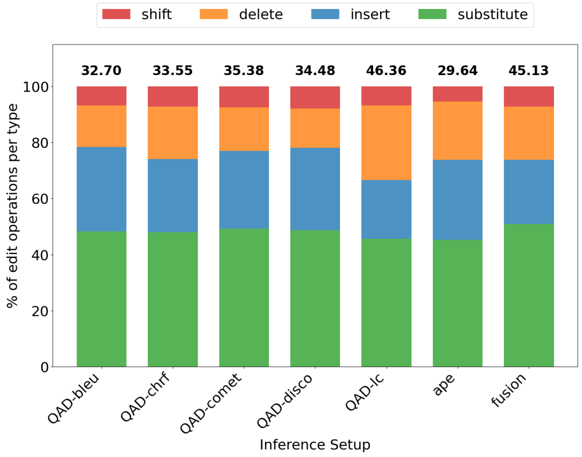

Figure 4: Edit rate analysis of inference setups against greedy outputs. The figure shows the proportion of each edit type as segments within the bar. The numbers on top represent the overall edit rate. The legend items, listed from left to right, correspond to the bar segments from top to bottom.

5.2 Human Qualitative Analysis

We conduct a small-scale manual qualitative analysis to better understand the impact of QAD on translation quality. This analysis helps us examine the semantic differences between greedy and QAD outputs. Additionally, it allows us to confirm findings from 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 samples 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 the translations. We mask the MT type information[5] and ask them to annotate the data, as follows (full guidelines in Appendix D):

- Identify any of the four linguistic phenomena in the source sentence and upon identification
    - ∗ Identify whether the phenomenon is translated correctly in (a) the greedy and (b) the QAD translation.
- Annotate the semantic difference between greedy and QAD hypotheses on a Likert scale of 1–5.
- Select their preference between the greedy and QAD hypothesis.

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

- Optionally, comment on their preference and observations.

Annotation statistics are presented in Table 6, 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 similarity against preferences are presented in Figure 3 (Arabic is excluded from these figures to remove model-specific bias; its results in Appendix B 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.

6 Related Work

Linguistic analysis of LLMs. Manakhimova et al. (2024) develop a fine-grained test suite to evaluate the linguistic performance of LLMs in MT, finding NMT systems outperform LLMs in phenomena like punctuation, future verb tenses, stripping, function words, etc. Sizov et al. (2024) highlight differences in lexical features between human, LLM, and NMT translations, showing LLMs align more closely with human translations in adverbs and auxiliary verbs, while NMT systems differ significantly. Raunak et al. (2023) find LLM translations are less literal than NMT translations but maintain equal or better quality. We extend these analyses by focusing on discourse phenomena and proposing the use of quality-aware decoding (QAD) to enhance discourse performance.

LLMs for Document-level translation. Wu et al. (2024) analyze LLMs tailored for document-level translation, examining translation errors, pronoun resolution, training and inference strategies, data efficiency of parallel documents, and zero-shot cross-lingual transfer. Efforts to adapt LLMs for document-level translation include finetuning the models using mixed sentence-level and document-level instructions (Li et al., 2024), prompting the models via in-context learning (Cui et al., 2024), and hybrid techniques that combine sentence-level translation models and monolingual document-level language models (Petrick et al., 2023). Unlike prior studies, we hypothesize that LLMs encode discourse knowledge and demonstrate that quality-aware decoding can effectively extract this knowledge, enabling LLMs to surpass encoder-decoder models in document-level translation tasks.

Gender bias in translation LLMs. As some phenomena we study can be affected by gender bias in the tested models, we present relevant works on gender bias in translation. Gender accuracy in translation can impact output fluency, translation accuracy, and ethics. Research efforts include creating challenging datasets (Currey et al., 2022; Rarrick et al., 2023; Jourdan et al., 2025), analyzing LLMs' performance (Zhao et al., 2024; Sánchez et al., 2024), identifying gender bias patterns in (Kotek et al., 2023) and mitigating it (Gupta et al., 2022; Sant et al., 2024).

7 Conclusion

We investigate the discourse phenomena performance of LLMs in document level translation. Specifically, we examine four aspects of discourse: lexical cohesion, formality, pronoun resolution, and verb forms. Our findings reveal that LLMs lag behind neural machine translation (NMT) systems 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 directions include exploring the use of this annotated data as a reward signal for fine-tuning LLMs to further enhance their discourse phenomena performance.

Limitations

- We rely on MuDA (Fernandes et al., 2023) for automatic tagging of discourse phenomena, and the tagging quality affects the discourse phenomena we are able to analyze. We use its default alignment and coreference resolution models, which may not represent the state of the art. Improving these components with better models could enhance tagging quality.

- We experiment with only one sampling approach (nucleus sampling); future work could investigate the impact of different sampling strategies 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 detailed experiments, supervision of models, etc. However, we leave it to future research as it is beyond the scope of this work.

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.

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.

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.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. 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.

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.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

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.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. Fairtranslate: An english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. *arXiv preprint arXiv:2504.15941*.

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.

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.

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

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, 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.

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.

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.

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.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. GATE: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pages 845–854. ACM.

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.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. 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.

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).

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.

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. 2024. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

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*.

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.

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.

A WMT24++ Results

Table 7 shows the results on WMT24 ++ for all language pairs. It is worth mentioning that all models exhibit low performance on EN-DE data. A manual qualitative analysis of the translations reveals that the reference translations are of suboptimal quality, often consisting of short sentences. LLM-as-a-judge scores are shown in Figure 5.

B Results of Human Qualitative Analysis on Arabic

Figure 6 shows the human annotations of the performance of QAD and greedy outputs on Arabic. Figure 7 shows the preference and semantic difference relationship for Arabic.

C Prompt Formats

Figures 8 and 9 present the prompt formats used to prompt the models.

D Human Assessment Details

Details of the data and instructions given to the annotators are presented in Table 8.

E 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).

Language Pair	Metric	**NLLB-3.3B**		**TowerInstruct-13B**		**EuroLLM-9B-Inst**	
		Greedy	QAD	Greedy	QAD	Greedy	QAD
EN-PT	BLEU	33.2	35.2	25.8	35.6	26.4	**39.3**
	COMET	78.8	79.5	79.2	**83.2**	78.1	**83.2**
	COMETQE	75.7	76.5	73.7	78.4	73.5	**79.1**
	Lexical Cohesion	77.0	76.0	78.0	83.0	79.0	**84.0**
	Formality	58.0	62.0	47.0	61.0	58.0	**66.0**
	Pronouns	49.0	50.0	42.0	49.0	46.0	**56.0**
EN-DE	BLEU	5.1	5.2	7.9	**12.6**	3.6	4.8
	COMET	48.1	47.5	58.2	**63.0**	48.2	50.8
	COMETQE	**77.1**	76.9	58.6	65.2	67.4	75.9
	Lexical Cohesion	26.0	27.0	22.0	23.0	**30.0**	29.0
	Formality	23.0	23.0	36.0	**37.0**	24.0	23.0
	Pronouns	**26.0**	23.0	22.0	14.0	**26.0**	23.0
EN-FR	BLEU	32.8	33.7	25.7	**36.9**	23.3	32.8
	COMET	75.8	75.7	76.6	**81.3**	74.0	79.7
	COMETQE	78.7	78.7	75.5	**81.1**	74.0	80.1
	Lexical Cohesion	70.0	70.0	72.0	**80.0**	66.0	72.0
	Formality	56.0	58.0	57.0	**61.0**	49.0	58.0
	Pronouns	49.0	46.0	50.0	**60.0**	47.0	53.0
	Verb form	37.0	**49.0**	30.0	45.0	37.0	**49.0**
EN-RU	BLEU	20.6	20.7	13.7	23.2	15.1	**23.5**
	COMET	76.1	76.2	73.8	81.5	76.0	**81.6**
	COMETQE	75.8	75.5	67.5	77.5	72.2	**78.8**
	Lexical Cohesion	63.0	53.0	65.0	72.0	64.0	**75.0**
	Formality	47.0	48.0	34.0	**55.0**	49.0	52.0
	Verb form	32.0	34.0	21.0	**38.0**	28.0	37.0
EN-AR	BLEU	17.5	16.9	N/A	N/A	10.3	**20.6**
	COMET	77.8	77.1	N/A	N/A	75.3	**81.7**
	COMETQE	72.7	71.0	N/A	N/A	66.8	**75.1**
	Lexical Cohesion	63.0	60.0	N/A	N/A	58.0	**75.0**
	Pronouns	**55.0**	43.0	N/A	N/A	36.0	54.0
EN-KO	BLEU	22.2	21.5	12.0	20.1	17.0	**28.6**
	COMET	80.1	79.2	77.4	83.4	79.9	**85.7**
	COMETQE	79.4	78.3	69.5	77.9	75.4	**82.1**
	Lexical Cohesion	44.0	32.0	34.0	43.0	38.0	**57.0**
	Formality	27.0	29.0	25.0	32.0	09.0	**34.0**

Table 7: Translation and discourse phenomena performance of the three models using greedy and QAD setups on WMT24++ dataset. N/A: not applicable as TowerInstruct-13B is not trained on Arabic. **Bold** highlights the best value per row. The results demonstrate that QAD enhances the performance of LLMs.

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 cohesion**: 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 8: Human assessment details.

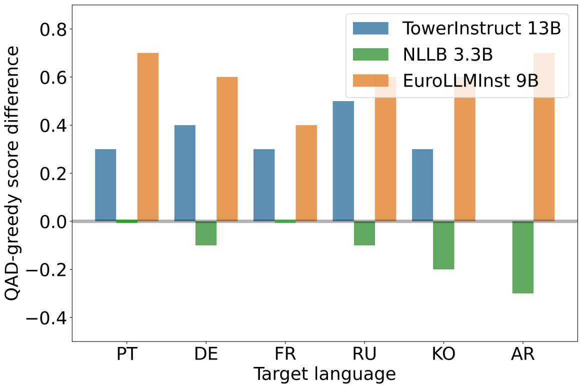

Figure 5: Difference between QAD and greedy LLM-as-a-judge scores on WMT24++ data. The plot demonstrates that QAD improves the performance of LLMs.

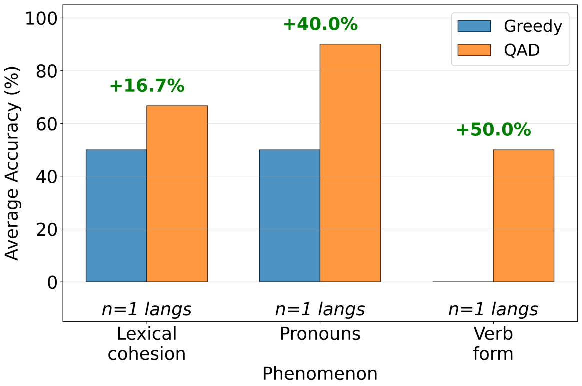

Figure 6: Human-annotated accuracy of greedy and QAD outputs in handling discourse phenomena for Arabic data.

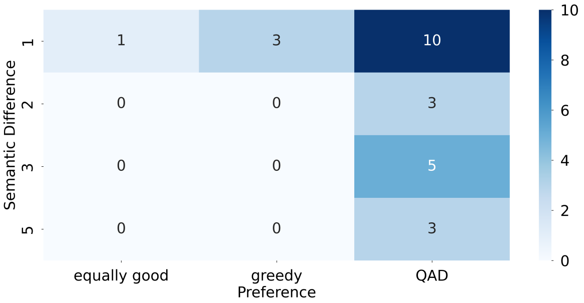

Figure 7: Semantic difference vs. preference on Arabic data.

```
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 8: TowerInstruct-13B 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>:
<src_lang>: <src sentence> <tgt_lang>:
```

Figure 9: EuroLLM-9B-Inst prompt format