Context-Aware or Context-Insensitive? Assessing LLMs' Performance in Document-Level Translation

Wafaa Mohammed **Vlad Niculae**
Language Technology Lab
University of Amsterdam
{w.m.a.mohammed, v.niculae}@uva.nl

Abstract

Large language models (LLMs) are increasingly strong contenders in machine translation. In this work, we focus on document-level translation, where some words cannot be translated without context from outside the sentence. Specifically, we investigate the ability of prominent LLMs to utilize the document context during translation through a perturbation analysis (analyzing models' robustness to perturbed and randomized document context) and an attribution analysis (examining the contribution of relevant context to the translation). We conduct an extensive evaluation across nine LLMs from diverse model families and training paradigms, including translation-specialized LLMs, alongside two encoder-decoder transformer baselines. We find that LLMs' improved document-translation performance compared to encoder-decoder models is not reflected in pronoun translation performance. Our analysis highlight the need for context-aware finetuning of LLMs with a focus on relevant parts of the context to improve their reliability for document-level translation.

1 Introduction

Language normally consists of collocated, structured, coherent groups of sentences referred to as a discourse (Jurafsky and Martin, 2009, chapter 21). Discourse properties that go beyond an individual sentence include the frequency and distribution of words within a document, topical, functional and discourse coherence patterns, and the use of reduced expressions. These properties have stimulated a good deal of machine translation research in the 1990s, aimed at endowing machine–translated target texts with the same properties as their source texts (Nash-Webber et al., 2013). Since then, there has been a growing interest in document-level translation, mainly focused on document-level influences on lexical choice, and developing methods, annotated resources and assessment metrics for discourse-level machine translation (Popescu-Belis et al., 2019).

Large language models (LLMs) show promise on multiple language technologies, with recent models specially finetuned for machine translation (Alves et al., 2024; Xu et al., 2023). Wang et al. (2023) suggest that translation LLMs have potential to be the new paradigm for document-level translation. While such work focuses only on assessing translation quality using metrics such as BLEU or COMET, our work investigates how models utilize context in translation. Inspired by Mohammed and Niculae (2024), we follow an interpretable approach towards context utilization evaluation. In particular, we focus on answering two main questions: how sensitive LLMs are to the correct context, and how well they utilize the relevant parts of context.

For context sensitivity assessment, we compare the general and discourse-phenomena-specific (Müller et al., 2018) translation performance of LLMs under the gold context setup to a perturbed context setup. For relevant-context utilization assessment, we perform a finer-grained evaluation. We look at models' internals using attribution methods (Ferrando et al., 2023) to quantify the contribution of relevant context to the translation. Context utilization in machine translation has been explored in encoder-decoder models, such as by Sarti et al. (2023), who developed an end-to-end interpretability framework to assess context-aware translation. To the best of our knowledge, we are the first to explore context utilization in translation LLMs via perturbation and attribution methods.

Our main findings can be summarized in the following:

Proceedings of Machine Translation Summit XX Volume 1, pages 126–137, June 23-27, 2025

- Translation-finetuned LLMs outperform encoder-decoder models at overall translation, but perform worse on discourse phenomena.

- Despite being smaller and not specifically finetuned for translation tasks, the EuroLLM-9B-Inst multilingual model outperforms the TowerInstruct 13B model at translation.

- All evaluated models show robustness to randomized context. We attribute this to lack of proper context utilization and highlight the need for explicit context-aware finetuning of LLMs to ensure their reliability for document-level translation.

- Our analysis of model internals reveals low *relevant-context* attribution scores, further highlighting the necessity for explicit context-aware finetuning.

The structure of our paper is as follows: §2 provides an overview of the analyses conducted, while §3 outlines the experimental setup. In §4, we present and discuss the results of our experiments. A review of additional related work is included in §5, and we present our conclusions and suggestions for future work in §6. Finally, §7 addresses the limitations of our research and our ethical considerations are detailed in §8.

2 Analysis overview

This section presents an overview of the analyses we conducted. Like Mohammed and Niculae (2024), we perform a perturbation analysis on translation quality and pronoun resolution accuracy. Moreover, we examine model mechanics through an attribution analysis via interpretability methods.

2.1 Perturbation Analysis

Translation quality. To assess model's sensitivity to gold context, we compare models' translation behavior in different context setups: a gold, perturbed, and random context setup on IWSLT2017 data (Cettolo et al., 2012). The gold context (Figure 1a) is the previous source-target pairs. For the perturbed context (Figure 1b), we randomly sample sentences from a different document, matching the size of the gold context. We sample sentences from a different document instead of the same document to ensure a robust analysis of models' reliance on relevant contextual information and to avoid introducing unintended biases due to implicit thematic or lexical similarities. Random context (Figure 1c) is uniformly-sampled random tokens from the model's vocabulary, with the same size as the gold context.

Pronoun resolution. We perform a phenomenon-specific assessment of models' sensitivity to gold context by comparing pronoun resolution performance in different context setups on ContraPro data (Müller et al., 2018; Lopes et al., 2020). We focus on pronoun resolution as a measurable phenomenon where perturbation experiments can be defined due to the availability of datasets with supporting context annotations. The gold and random contexts (Figures 2a and 2c) are the same as for IWSLT2017 data. Here, instead of the perturbed context replacing the gold context with sentences from different documents, we only replace antecedent tokens in the gold context with different-gender tokens (Figure 2b). This allows for a finer-grained context-utilization analysis. We create a database of antecedent words, clustered by POS (Part Of Speech) tag and gender. Each antecedent is replaced with a random word of the same POS tag but different gender. For antecedents with rare POS tags (0.2% of cases), no such alternative can be found, so we sample a random different-gender word with any tag.

2.2 Attribution Analysis

For a finer-grained evaluation, we analyze how much LLMs utilize relevant context when translating ambiguous pronouns. We use two existing attribution methods: ALTI-Logit (Ferrando et al., 2023) and input-erasure (Li et al., 2016), as Krishna et al. (2022) point out that state-of-the-art explanation methods often disagree. ALTI-Logit tracks the logit (pre-activation of the softmax) contributions back to the input by aggregating across layers and considering the mixing of information in intermediate layers using ALTI (Ferrando et al., 2022). Input-erasure measures the change in model's prediction when removing parts of the input. Attribution methods provide for every token in the model input X, a non-negative attribution score $\{a_t : t \in X\}$, corresponding to the amount that token contributes to the next token prediction. For our aim, we measure how much of the overall attribution goes to a subset of the input $S \subseteq X$. This motivates the

attribution percentage:

$$\text{AP}(S)\% = \frac{\sum_{t \in S} a_t}{\sum_{t \in X} a_t} \times 100\%. \quad (1)$$

3 Experimental Details

This section includes details about models, datasets, prompt formats, and evaluation metrics used in our experiments. The sustainability statement for our experiments is presented in Appendix A.

3.1 Models

We experiment on three model categories to capture the effects of large scale training, multilingual pretraining, and translation-specific finetuning.

Translation-finetuned LLMs. From the Tower family (Alves et al., 2024) we consider TowerBase, built on top of Llama-2 by continuing pretraining on multilingual data, and TowerInstruct which further finetunes TowerBase for translation-related tasks. We also analyze ALMA (Xu et al., 2023), which follows a two-step finetuning approach also on top of Llama-2, with multilingual and parallel data. As the foundation of the models above, we also include Llama-2 (Touvron et al., 2023), in order to capture the effects of translation-specific finetuning on context use. We consider the 7B and 13B versions of all models wherever feasible.

Multilingual LLMs. We experiment on EuroLLM-9B-Inst (Martins et al., 2024), a model trained on 35 languages, encompassing all European Union languages and additional relevant ones. Specifically, we use the instruction-tuned version of EuroLLM-9B-Inst to evaluate the impact of (multilingual pretraining + instruction tuning) compared to the (monolingual pretraining + continued multilingual pretraining + translation-specific fine-tuning) of Tower models.

Encoder-decoder baselines. We analyze NLLB-3.3B (Costa-jussà et al., 2022) as one of the state-of-the-art encoder-decoder translation models. As NLLB is trained at the sentence-level and not intended for document-level translation, we include only its sentence-level scores. As a context-aware encoder-decoder baseline, we also include a transformer-small model trained on the training subset of IWSLT2017 TED data (Cettolo et al., 2012). In specific, we train a small encoder-decoder transformer model (Vaswani et al., 2017) (hidden size of 512, feedforward size of 1024, 6 layers, 8 attention heads). We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and use an inverse square root learning rate scheduler with an initial value of 5×10^{-4} and with a linear warm-up in the first 4000 steps. We train the model with early stopping on the validation perplexity. The model is trained using a dynamic context size of 0–5 previous source and target sentences to ensure robustness against varying context size, as recommended by Sun et al. (2022). The training is performed on top of Fairseq (Ott et al., 2019).

3.2 Datasets

General translation assessment data. We evaluate on IWSLT2017 TED data (Cettolo et al., 2012), in English to German (EN→DE) and English to French (EN→FR). For EN→DE, we combine `tst2016-2017` for a test set of 2,271 sentences across 23 documents. For EN→FR, we use `tst2015`, containing 1,210 sentences in 12 documents. Following Mohammed and Niculae (2024), we use a context size of 5 previous source-target pairs. Future work could investigate the impact of context size on translation performance.

Pronoun resolution experiments data. We use ContraPro, a subset of OpenSubtitles (Müller et al., 2018; Lopes et al., 2020), consisting of examples with ambiguous pronouns, their gold translations, and automatic annotation of antecedents (relevant context) needed for resolution. For EN→DE, the dataset considers the translation of the English pronoun “it” to the three German pronouns “er”, “sie” or “es”. For EN→FR, the dataset concerns the translation of the English pronouns “it”, “they” to their French correspondents “il”, “elle”, “ils”, and “elles”. The dataset is balanced and consists is 12K instances for EN→DE and 14K instances for EN→FR. Our experiment is controlled: we experiment on instances where the antecedent distance is in the interval [1,5] in sentences and use 5 source-target pairs as context at inference time.

Attribution analysis data. Using ContraPro, we force-decode up to the pronoun, and measure the attribution percentage of the entire context and the relevant context (antecedents). Due to computational constraints, we analyze only the 7B version of LLMs in addition to EuroLLM-9B-Inst, randomly sample a balanced 2k subset of ContraPro and use a context size of 2.

```
English: When I was a kid, my parents would tell me, "You can make a mess, but you have to clean up after yourself."
German: Als Kind sagten mir meine Eltern immer: "Du kannst Unordnung machen, solange du hinterher aufräumst."
English: So freedom came with responsibility.
German: Freiheit war also mit Verantwortung verbunden.
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(a) Gold-context prompt

```
English: Before becoming a writer, Nora was a financial planner.
German: Bevor sie Autorin wurde, war Nora Finanzplanerin.
English: She had to learn the finer mechanics of sales when she was starting her practice, and this skill now helps her write compelling pitches to
↪ editors.
German: Sie befasste sich detailliert mit Verkaufsmechanismen, als sie ihre Praxis eröffnete. Diese Fertigkeit hilft ihr nun beim Entwickeln von
↪ Pitches für Redakteure.
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(b) Perturbed-context prompt

```
English: ro practicevalue downloadingcorezDescription Hence tierra Pur SeleAP hrefpick bore Engel delegate We WCF broad quattro bird stru corsategor
↪ ". nuc
German: Itemactivityrightarrow früher spend Universität Bull ^Password cantonmys@", largvarphikoamiltonounrenceoking říavctor NickFoot Colors
↪ stoneitosweh epe limits translate
English: ctoo Ski| anth https Baby Platform
German: HERannel/*medialabelignonliteretzt media Mittłurown
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(c) Random-context prompt

Figure 1: The figure shows example prompts used in the perturbation experiments for translation quality analysis, the reference translation (the last line of each example) is shown in green. The examples shown employ the explicit prompt format.

```
English: One of the Chinese worked in an amusement park.
German: Ein Chinese arbeitete in einem Vergnügungspark.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(a) Gold-context prompt

```
English: One of the Chinese worked in an house.
German: Ein Chinese arbeitete in einem Haus.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(b) Perturbed-context prompt

```
English: ro practicevalue downloadingcorezDescription Hence tierra Pur SeleAP hrefpick bore.
German: Itemactivityrightarrow früher spend Universität Bull ^Password.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(c) Random-context prompt

Figure 2: The figure shows example prompts used in the perturbation experiments for pronoun resolution analysis, the reference translation (the last line of each example) is shown in green. The pronoun of interest and its antecedents are highlighted in underlined blue. The examples shown employ the generic prompt format.

3.3 Evaluation

We evaluate translations using BLEU (Papineni et al., 2002), CHRF (Popović, 2015), and COMET (Rei et al., 2022). We also measure and pronoun translation accuracy in a contrastive force-decoded setting (CPRO; Müller et al., 2018) and a generative one (GPRO; Post and Junczys-Dowmunt, 2023). The contrastive pronoun resolution metric (CPRO) evaluates the models' accuracy in assigning a higher score to a sentence containing the correct pronoun compared to sentences with incorrect pronouns. The generative pronoun resolution metric (GPRO) assesses models' accuracy in generating the correct pronoun during inference. As Post (2018) points out the importance of providing SacreBLEU signatures for reproducibility, the details of our metrics are in Table 1.

metric	signature
BLEU	nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.4.0
CHRF	nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.0
COMET	https://huggingface.co/Unbabel/wmt22-comet-da

Table 1: Evaluation-metrics signatures

3.4 Prompt Format

Wu et al. (2024) noted that prompt formats significantly impact LLMs' performance, with well-structured prompts boosting models' performance. We use 3 formats from their work as in Fig. 3.[1]

4 Results and Discussion

This section presents and discusses the experimental results, covering the performance under the gold

[1]For TowerInstruct, we add an instruction-following prefix as per its documentation:<|im_start|>user {**prompt**} <|im_start|>assistant.

	Sentence baseline		Generic prompt random		Generic prompt perturbed		Generic prompt gold		Explicit prompt random		Explicit prompt perturbed		Explicit prompt gold	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
EN→DE														
Concat Enc-Dec	75.4	23.4	67.9	20.2	75.3	23.4	75.4	23.6	–	–	–	–	–	–
NLLB 3.3B	84.4	28.2	–	–	–	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	85.8	28.6	85.2	27.9	85.7	28.8	86.3	**30.8	85.4	28.3	85.7	28.8	**86.4	30.3
Llama-2 7B	79.0	20.8	42.6	01.5	79.8	21.3	81.2	22.0	77.9	20.1	79.8	21.6	81.2	22.8
Llama-2 13B	76.0	02.1	56.8	06.0	81.6	23.2	82.8	25.5	78.4	22.5	67.0	02.2	76.4	01.7
TowerBase 7B	82.8	25.8	82.1	25.7	83.7	25.9	83.8	25.6	83.0	26.3	82.5	26.4	82.0	26.3
TowerBase 13B	82.7	27.1	83.5	27.3	84.2	27.8	85.0	28.9	83.4	27.2	74.9	23.9	78.2	25.8
ALMA 7B	82.9	24.8	77.1	15.7	82.3	23.0	83.4	25.3	82.4	23.4	82.7	22.7	83.7	24.5
ALMA 13B	83.8	26.2	73.7	17.3	83.2	24.9	84.3	27.1	73.7	25.6	83.6	25.6	83.4	27.1
TowerInstruct 7B	84.8	27.3	84.4	26.6	84.8	27.0	85.2	27.5	84.4	26.4	84.7	27.0	85.0	27.1
TowerInstruct 13B	85.1	28.4	84.8	27.2	85.2	28.0	85.6	29.1	84.9	27.5	85.1	27.8	85.4	28.6
EN→FR														
Concat Enc-Dec	77.8	35.8	68.2	28.9	77.3	35.4	77.5	36.0	–	–	–	–	–	–
NLLB 3.3B	84.8	38.5	–	–	–	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	86.4	40.8	85.9	40.3	86.5	41.3	**86.8	**43.4	86.2	40.5	86.3	41.4	86.7	42.8
Llama-2 7B	81.6	33.2	29.5	01.2	81.8	29.6	82.6	34.7	80.9	31.6	82.0	31.5	82.5	30.9
Llama-2 13B	77.0	17.1	54.7	04.2	83.8	35.5	84.5	38.4	81.1	34.2	81.9	20.7	83.4	06.3
TowerBase 7B	84.7	39.9	83.8	37.1	79.0	10.8	78.7	36.2	84.4	40.0	79.1	13.6	76.5	35.4
TowerBase 13B	79.4	39.5	84.9	41.0	85.1	40.7	85.9	41.9	85.1	40.7	85.4	40.6	69.3	31.7
ALMA 7B	80.8	28.7	52.2	07.1	80.4	25.7	81.1	27.9	80.3	28.9	80.5	27.4	81.3	30.5
ALMA 13B	83.0	33.7	60.0	10.0	82.8	32.7	83.4	33.1	82.9	33.9	82.9	33.9	83.7	35.1
TowerInstruct 7B	85.8	38.1	85.5	35.4	83.4	33.0	86.0	39.6	85.4	36.1	84.1	36.9	85.9	39.1
TowerInstruct 13B	86.2	40.0	86.0	39.3	86.0	40.3	86.4	40.9	86.0	39.5	86.0	40.2	86.2	40.8

Table 2: Translation performance (COMET and BLEU) on IWSLT2017, with random, structurally perturbed and gold context, for the prompts considered. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

	sentence			random			perturbed			gold		
	COMET	GPRO	CPRO	COMET	GPRO	CPRO	COMET	GPRO	CPRO	COMET	GPRO	CPRO
EN→DE												
Concat Enc-Dec	66.2	41.7	46.4	61.5	32.6	45.3	66.9	53.5	**60.4	67.0	**56.2	**60.4
NLLB 3.3B	**72.3	41.6	32.0	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	61.5	29.7	54.7	50.9	24.5	51.0	41.6	21.8	47.7	43.7	29.6	51.4
Llama-2 7B	35.0	09.7	45.2	27.6	02.3	46.3	39.3	22.1	46.9	41.6	26.1	49.9
Llama-2 13B	34.2	07.6	45.1	28.0	03.0	45.9	40.1	25.5	49.6	42.7	31.1	56.7
TowerBase 7B	39.6	14.1	46.7	35.0	11.2	45.7	44.0	25.1	47.9	45.9	28.9	50.8
TowerBase 13B	56.6	30.8	46.6	31.8	06.6	46.4	51.6	27.3	49.9	50.2	32.2	53.8
ALMA 7B	52.4	22.1	46.4	30.7	06.8	45.8	46.5	25.6	47.2	49.0	30.6	49.9
ALMA 13B	55.3	24.6	46.9	30.3	05.7	47.5	46.3	29.7	52.2	48.6	35.5	58.5
TowerInstruct 7B	57.0	29.9	49.8	40.7	14.5	58.0	53.9	27.1	48.5	55.2	30.7	51.9
TowerInstruct 13B	56.6	30.8	54.5	53.8	21.8	59.2	51.6	27.8	55.0	60.9	32.2	59.9
EN→FR												
Concat Enc-Dec	66.5	51.7	76.5	62.7	51.6	76.2	66.8	57.7	80.5	67.0	**65.0	86.0
NLLB 3.3B	**76.3	64.0	36.9	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	58.5	34.2	06.7	28.8	00.7	17.0	43.2	25.4	11.6	46.9	36.7	13.2
Llama-2 7B	38.0	12.9	90.0	28.7	01.5	64.6	41.9	24.8	64.5	46.1	34.0	68.2
Llama-2 13B	34.1	6.3	89.4	29.1	02.2	49.0	42.5	25.6	59.2	47.1	35.1	63.6
TowerBase 7B	41.5	14.7	**94.5	38.5	09.8	70.2	45.7	26.7	85.9	50.2	36.3	88.1
TowerBase 13B	38.0	10.1	78.3	33.7	05.7	74.3	47.6	28.4	80.1	52.5	38.3	82.1
ALMA 7B	42.6	14.7	11.2	29.1	02.4	05.4	41.7	22.7	09.0	45.4	29.7	10.6
ALMA 13B	45.0	16.5	09.4	30.1	03.0	05.3	44.4	26.7	08.3	48.6	34.4	09.8
TowerInstruct 7B	56.6	35.9	55.1	34.9	04.0	23.8	50.3	29.3	52.6	55.1	39.5	56.5
TowerInstruct 13B	57.0	35.1	11.1	47.9	14.1	04.7	53.1	30.3	12.4	58.1	40.4	13.8

Table 3: This table presents the translation performance measured using COMET, the generative (GPRO) and the contrastive (CPRO) pronoun-resolution accuracies on ContraPro dataset, with random, structurally perturbed and gold context, and generic prompt. Random guessing accuracy: 33.3% EN→DE, 50% EN→FR. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

```
Translate the following <src_lang> source text to <tgt_lang>:     (a)
<src_lang>: <src_sentence> <tgt_lang>:
```

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>           (b)
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
<src_lang>: <src sentence> <tgt_lang>:
```

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>           (c)
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
Given the provided parallel sentence pairs, translate the following
↪  <src_lang> sentence to <tgt_lang>:
<src_lang>: <src sentence> <tgt_lang>:
```

Figure 3: a) sentence-level, b) generic, and c) explicit prompt formats. `tgt context` refers to gold translations.

context setup, the perturbation analysis (performance under the perturbed and random context setups), and the attribution analysis looking at the models' internals.

4.1 Performance With the Gold Context

Overall translation performance. Table 2 shows the translation performance (BLEU, COMET) on IWSLT2017 in the sentence-level baseline setup, the generic prompt setup, and the explicit prompt setups. CHRF results are in a separate table (Table 4) for better readability. We analyze the results of different model categories and summarize the observations and their intuitions in the following paragraphs.

We notice that document-level *generic* prompting improves translation performance of all models over the sentence-level baseline. This is expected since document-level prompting gives the model access to inter-sentential context. Moreover, *explicit* prompting improves instruction-finetuned models' performance, while strong base-models (such as TowerBase 13B) degrade in performance. This is also aligned with expectations of the sensitivity of models to the prompt format (Wu et al., 2024), and it highlights the importance of aligning training and inference prompts. However, as the gains with explicit prompting are not substantial even for instruction-tuned models, we opt for the generic prompt format for the pronoun resolution experiments.

For models under consideration in this work, decoder-only LLMs outperform encoder-decoder models at overall translation. This aligns with previous research findings of the potential of LLMs as a new paradigm for document-level translation (Wang et al., 2023). Interestingly, for both language pairs, EuroLLM-9B-Inst outperforms all models in both prompting formats. In the explicit prompting format, TowerInstruct 13B achieves the second-highest performance, while in the generic format, TowerBase 13B comes in second (for EN→FR). EuroLLM-9B-Inst's recipe of multilingual pretraining and instruction tuning seems to have better effects on improving the translation performance compared to the continued multilingual pretraining and translation-specific fine-tuning of Tower models. ALMA models lag behind Tower models despite both employing a two-step fine-tuning strategy on multilingual and parallel data. This raises the need for a deeper investigation into how various design choices (such as the selection and number of finetuning languages, the choice of datasets, and the configuration of hyper-parameters) influence downstream performance.

Further analyzing Table 2, we observe that Llama-2 13B model has a noticeably low performance with explicit gold context for both language pairs. While surprising at first sight, we argue that as the model is pretrained mainly on English data, it might not be sufficient for this task. We look at the translations produced by the model and find that they are mostly repeated words or outputs in the source language instead of the target language.

Pronoun resolution performance. Table 3 shows the generative and contrastive pronoun accuracy and translation performance (COMET) on ContraPro dataset.

Similar to the overall translation performance, We notice that document-level prompting outperforms sentence-level prompting in pronoun resolution performance. A key finding from this analysis is the contrasting ranking compared to the overall translation performance: both encoder-decoder baselines outperform all LLMs in terms of GPRO and COMET scores. Even with gold context, LLMs' performance remains notably poor, with accuracy at or below the random guessing accuracy (33.3% for EN→DE, and 50% for EN→FR). This suggests that there is room to improve LLMs' translation finetuning to better handle context-dependent discourse phenomena.

However, it is important to note that except for the encoder-decoder transformer model that we trained from scratch, we don't have access to other models' training data, therefore, we cannot guarantee that ContraPro is unseen and thus that the evaluation is fair. In particular, NLLB's performance far above chance at the sentence level may be due to such contamination, as sentence-level evaluation forces it to *guess* the pronoun gender

	Sent. base.	rand.	Genric pert.	gold	rand.	Explicit pert.	gold
EN→DE							
Concat Enc-Dec	53.0	50.7	53.0	53.1	–	–	–
NLLB 3.3B	**59.7**	–	–	–	–	–	–
EuroLLM-9B-Inst	59.4	**58.8**	**59.1**	**60.4**	**59.2**	**59.5**	****60.7**
Llama-2 7B	51.2	12.1	51.3	52.2	51.0	52.0	53.3
Llama-2 13B	35.1	17.9	53.5	54.8	52.2	32.5	33.5
TowerBase 7B	56.9	56.7	57.0	56.4	57.1	56.8	56.5
TowerBase 13B	57.8	57.9	58.3	59.1	57.9	51.7	54.8
ALMA 7B	54.8	46.6	53.0	54.8	54.5	54.2	55.4
ALMA 13B	56.6	43.5	55.2	56.8	56.2	56.2	57.4
TowerInstruct 7B	57.9	57.4	57.7	58.1	57.4	57.7	57.9
TowerInstruct 13B	58.9	58.2	58.6	59.4	58.2	58.5	59.1
EN→FR							
Concat Enc-Dec	60.9	56.4	60.9	61.3	–	–	–
NLLB 3.3B	**65.9**	–	–	–	–	–	–
EuroLLM-9B-Inst	65.6	65.2	**66.0**	****67.4**	65.8	**66.4**	**67.3**
Llama-2 7B	59.1	06.5	58.3	60.0	59.0	59.2	59.4
Llama-2 13B	55.6	15.1	61.8	63.2	60.0	59.2	51.7
TowerBase 7B	65.5	64.6	44.2	58.9	65.5	48.4	58.5
TowerBase 13B	64.4	**66.2**	65.9	66.6	**66.0**	65.8	55.2
ALMA 7B	56.6	20.4	54.9	55.8	56.6	55.6	57.9
ALMA 13B	59.9	25.3	59.8	60.4	59.7	60.5	61.4
TowerInstruct 7B	64.2	63.0	62.8	65.2	63.3	64.3	64.9
TowerInstruct 13B	65.2	64.9	65.4	65.9	64.9	65.5	65.6

Table 4: CHRF scores on IWSLT2017 test data for the sentence-level baseline and the random, structurally perturbed and gold context, for the prompts considered. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

without antecedent information.

Contrastive evaluation measures the classification accuracy of models which does not necessarily correlate with the generative training objective. As suggested by Post and Junczys-Dowmunt (2023), generative scores are better at discriminating document-level systems compared to contrastive scores, which is what we notice in CPRO results where we see surprising trends, with TowerBase 7B leading in EN→FR and TowerInstruct 13B performing comparably to the Concat Enc-Dec model in EN→DE which doesn't align with their GPRO and COMET performance on the data.

4.2 Perturbation Analysis

Structurally perturbed context. From Table 2, we see that structurally perturbing the context has a minimal impact on overall translation performance. All models exhibit only a slight degradation in BLEU, COMET, and CHRF scores when provided with a perturbed context. However, a closer look at the impact of context perturbation on pronoun resolution performance (Table 3) reveals more pronounced effects. Specifically, there is a notable decrease in GPRO performance, ranging from -5 to -10 points, under perturbed context conditions. Nevertheless, the similar level of performance reduction across models suggests that no model stands out in its ability to leverage context effectively. This can be attributed to the fact that none of the models are explicitly trained for context utilization.

Random context. Looking at models' performance with total random tokens, we find that on IWSLT data, EuroLLM-9B-Inst and Tower models (the best at translation) are robust to random context and only degrade slightly in performance, aligning with previous observations of the minimal effect of context perturbation on translation performance. Additionally, those models (except EuroLLM-9B-Inst for EN→FR) show the least difference in GPRO performance between gold and random context setups among all LLMs. Robustness to total random context can be linked to lack of proper context utilization. Although the TowerBlocks dataset used to finetune TowerInstruct models includes context-aware data (as per the dataset card[2]), we hypothesize that general fine-tuning alone may not be sufficient for improving discourse phenomena performance. Explicit, context-aware fine-tuning might be required to ad-

[2]https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.2

dress these challenges effectively.

Further analyzing Table 2, it's noteworthy that the TowerBase 7B model performs better with random context as compared to gold context, even though the latter resembles a few-shot learning scenario (Reinauer et al., 2023). That said, we point out that its translation performance is suboptimal, as it is an intermediate model between the base model Llama-2 7B and the instruction-tuned model TowerInstruct 7B designed specifically for translation.

4.3 Attribution Analysis

We analyze models' internals to see how much the relevant context contributes to the outputs. Figures 4a and 4b present attribution percentages of antecedent tokens (relevant context) as well as of the whole context using ALTI-Logit and input-erasure methods, respectively.

Looking at both attribution methods, we see that for EuroLLM-9B-Inst and TowerInstruct 7B (the best two models at translation among the 5 models tested) antecedent tokens have the lowest attribution percentage to the output. Even though for the TowerInstruct 7B model, overall context tokens have the highest attribution percentage. This suggests that there is a need to explicitly finetune translation LLMs to focus on *relevant context* at inference time.

However, unlike the larger differences in *relevant context* and overall context attributions observed for encoder-decoder models by Mohammed and Niculae (2024), we find no striking differences or clear patterns between the contributions for LLMs. This might be due to the fact that the models have similar backbone structures.

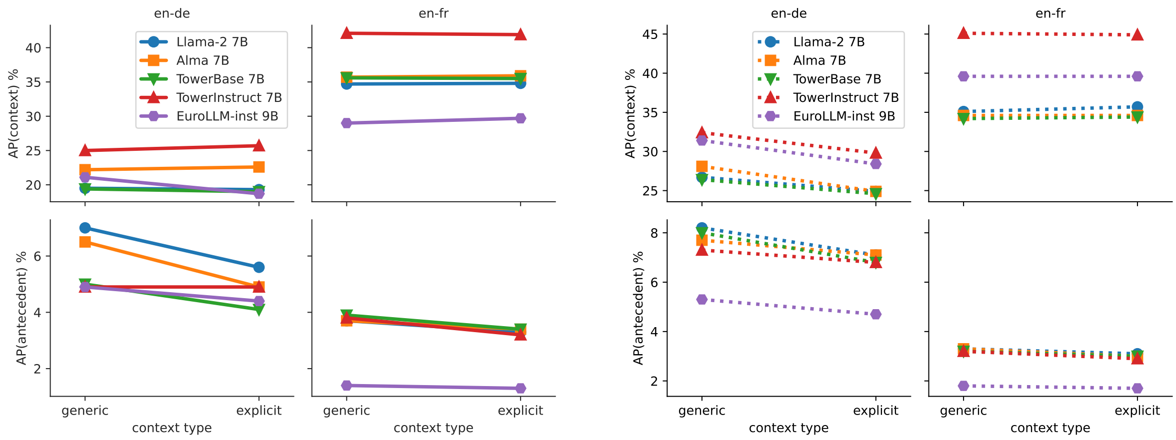
5 Related Work

Context utilization assessment. Works on assessing context utilization in machine translation include the work of Sarti et al. (2023), who build an end-to-end interpretability framework to quantify the plausibility of context-aware encoder-decoder machine translation models. They leverage model internals to contrastively identify context-sensitive target tokens in generated texts and link them to contextual cues justifying their prediction. Using their approach, they were able to consistently detect context-sensitive tokens and their disambiguating rationales across several datasets, models and discourse phenomena. Inspired by this line of research, we evaluate context utilization of LLMs as a possible new paradigm for context-aware translation.

Perturbation and attribution analysis. There are several works that used attribution and perturbation techniques to understand the inner workings of translation LLMs, mostly focusing on the in-context learning (ICL) paradigm —a setup where LLMs "learn" to perform new tasks during inference by being provided with few task demonstrations in the input prompt. Zaranis et al. (2024) use input attribution methods (ALTI) to examine context contributions in translation LLMs within the ICL paradigm. Their findings indicate that the source segments of few-shot examples contribute more significantly than their corresponding target segments, parallel-data fine-tuning alters contribution patterns, and context contributions exhibit a positional bias. Raunak et al. (2023) perturb in-domain translations to better understand their role in ICL. They perform asymmetric perturbation of source-target mappings and find that target perturbations has more negative effect on the translation performance compared to source perturbations. Zhu et al. (2024) also perturb the in-context examples by providing unrelated task (summarization) examples and find that LLMs are not sensitive to the perturbation. Our work combines both interpretability techniques (perturbation and attribution methods) and focuses on context-aware translation task.

LLMs for document-level machine translation. The line of research on adapting LLMs for document-level translation using techniques like LLMs fusion with translation models (Petrick et al., 2023), finetuning LLMs on parallel document-level data (Wu et al., 2024), or a mix of sentence-level and document-level data (Li et al., 2024), generally evaluates on translation metrics and discourse phenomenon accuracy. We complement such evaluations with a finer grained strategy that focuses on the role of context.

Gender bias. Although gender bias does not directly impact our analysis of pronoun resolution —, given that the referents in the ContraPro data are common nouns with clear grammatical gender and, in most cases (the entire German dataset and at least half of the French dataset), are non-human—we recognize that gender bias remains a significant



(a) ALTI-Logit

(b) input-erasure

Figure 4: Attribution percentages assigned to antecedent tokens (relevant context) and the entire context tokens when force-decoding the correct pronoun in ContraPro data. (a) shows results from ALTI-Logit and (b) displays results from input-erasure attribution methods.

concern for machine translation models and LLMs, as widely explored in research (Rudinger et al., 2018; Zhao et al., 2018; Currey et al., 2022; Rarrick et al., 2023)

6 Conclusion

We use interpretability tools (perturbation and attribution techniques) to analyze LLMs' context-utilization in document-level translation. Our experiments suggest that multilingual pretraining and translation-specific finetuning of LLMs pushes state-of-the-art translation performance beyond encoder-decoder models. However, we highlight that looking at discourse phenomena performance, LLMs show room for improvement. We argue that more care is needed before adopting LLMs as the new standard for document-level translation, and more focused evaluation must be considered. Future research directions include enhancing context-aware translation capabilities of LLMs, potentially through explicit finetuning, and creating datasets with supporting-context annotations for other discourse phenomena to enable extending context-utilization analysis to those phenomena.

7 Limitations

Even though some API-only LLMs (GPT-3.5 and GPT-4) show significant translation improvement compared to encoder-decoder document-level transformers and commercial translation systems (Wang et al., 2023), our analysis approach relies on access to model internals in order to be able to compute attributions of input tokens. Thus, we only used open-source LLMs in our study.

Based on the availability of datasets with context-dependent linguistic phenomena that include supporting context annotations (ContraPro), we experimented only on EN→DE and EN→FR. These two languages belong to the same language family; we leave it to future work to experiment on general translation on other language families.

We chose well-established evaluation metrics in the literature to assess pronoun resolution accuracy. However, we acknowledge the limitations of those metrics. The contrastive metric (CPRO) is not aligned with the generative training objective of models and the generative metric (GPRO) misses cases where the model generates the correct pronoun in a different location in the sentence than the target location.

Due to computational constraints, we were only able to perform the attribution analysis on a small set of models. We hope our work inspires more research into understanding the inner-workings of translation models in context utilization.

For all models except the transformer encoder-decoder model trained from scratch, we do not have details about their training data. This trend of releasing and building on models with secret training data is concerning because it makes fair evaluation impossible.

In our work, we focused on a fine-grained evaluation of context use on a specific phenomenon. Nonetheless, pretrained context-aware metrics could offer more accurate insights into overall models' performance on context use.

8 Ethics Statement

Nowadays, machine translation is a widely adopted technology, sometimes in sensitive, high-risk settings. Even though we propose a fine-grained approach to assessing context utilization, and highlight its importance as we see that improvements in translation performance does not necessarily reflect in discourse phenomena performance, we still rely on automatic evaluation which is imperfect. For systems deployed in critical scenarios, we believe a nuanced case-by-case evaluation is always necessary.

Acknowledgements

We would like to thank Evgenia Ilia, Bryan Eikema and Di Wu for their valuable comments and discussions about this work.

This work is part of the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631. VN also acknowledges support from the Dutch Research Council (NWO) via VI.Veni.212.228.

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation, EAMT 2012, Trento, Italy, May 28-30, 2012*, pages 261–268. European Association for Machine Translation.

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

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.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8698–8714. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5486–5513. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *CoRR*, abs/2202.01602.

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

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

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.

António V. 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, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, 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. *Preprint*, arXiv:2409.16235.

Wafaa Mohammed and Vlad Niculae. 2024. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1633–1643. 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics.

Bonnie Nash-Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation, DiscoMT@ACL 2013, Sofia, Bulgaria, August 9, 2013*. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. 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, WMT 2023, Singapore, December 6-7, 2023*, pages 375–391. Association for Computational Linguistics.

Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong, editors. 2019. *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

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*, pages 845–854.

Vikas Raunak, Hany Hassan Awadalla, and Arul Menezes. 2023. Dissecting in-context learning of translations in gpts. *CoRR*, abs/2310.15987.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriele Sarti, Grzegorz Chrupala, Malvina Nissim, and Arianna Bisazza. 2023. Quantifying the plausibility of context reliance in neural machine translation. *CoRR*, abs/2310.01188.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16646–16661. 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. *CoRR*, abs/2309.11674.

Emmanouil Zaranis, Nuno Guerreiro, and André F. T. Martins. 2024. Analyzing context contributions in llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14899–14924. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. 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, Mexico City, Mexico, June 16-21, 2024*, pages 2765–2781. Association for Computational Linguistics.

A Sustainability statement

Our experiments with 13B parameter models run in 95h on 2 GPUs NVIDIA A100 PCIe, and draw 81.69 kWh. Based in the Netherlands, this has a carbon footprint of 30.58 kg CO2e, which is equivalent to 2.78 tree-years. For all other models, the experiments run in 502h on 1 GPU NVIDIA A100 PCIe, and draw 222.08 kWh. Based in the Netherlands, this has a carbon footprint of 83.13 kg CO2e, which is equivalent to 7.56 tree-years (Lannelongue et al., 2021).