Analyzing Context Utilization of LLMs in Document-Level Translation

Anonymous ACL submission

Abstract

Large language models (LLM) are increasingly strong contenders in machine translation. We study document-level translation, where some words cannot be translated without context from outside the sentence. We investigate the ability of prominent LLMs to utilize context by analyzing models' robustness to perturbed and randomized document context. We find that LLMs' improved document-translation performance is not always reflected in pronoun translation performance. We 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). Document-level discourse properties (such as coherence patterns and frequency of words) have stimulated a good deal of machine translation (MT) 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 MT (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 on the document level as well. 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 investigate how sensitive LLMs are to the correct context, and how well they utilize the relevant parts of context.

For context utilization assessment, we compare general and discourse-phenomena-specific (Müller et al., 2018) translation performance with a perturbed context against the gold context. For 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 translation has been explored in encoder-decoder (enc-dec) models, such as by Sarti et al. (2023), who developed an 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. We find that translation-finetuned LLMs outperform enc-dec models at overall translation, but perform worse on discourse phenomena. Therefore, we highlight the need for context-aware finetuning of LLMs to improve discourse phenomena performance. Moreover, we point out that fine-grained evaluation of the internals of LLMs is necessary to understand their context utilization abilities.

2 Methodology

2.1 Analysis Overview

To assess model's sensitivity to gold context, we compare models' translation behavior in different context setups (examples in Appendix B). The **gold context** is previous source-target pairs. For the **perturbed context**, we randomly sample sentences from a different document, matching the gold context size. **Random context** is uniformly-sampled random tokens from the model's vocabulary, with the same size as the gold context.

We also perform a phenomenon-specific assessment of model's sensitivity to gold context. Here,

	Sentence baseline		Generic prompt						Explicit prompt					
			random		perturbed		gold		random		perturbed		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	–	–	–	–	–	–	–	–	–	–	–	–
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	–	–	–	–	–	–	–	–	–	–	–	–
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 1: Translation performance (COMET and BLEU) on IWSLT2017, with random, structurally perturbed and gold context, for the prompts considered. Bold marks the best value per column; (**) marks best overall.

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. 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.

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) points out that state-of-the-art explanation methods often disagree. ALTI-Logit tracks the logit 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**:

$$\mathrm{AP}(S)\% = \left(\sum_{t \in S} a_t\right) / \left(\sum_{t \in X} a_t\right) \times 100\%. \quad (1)$$

2.2 Models, Datasets and Evaluation

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 monolingual 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.

Encoder-decoder baselines. We include a transformer small model trained on IWSLT2017 with concatenated context (details in Appendix A) and NLLB-3.3B (Costa-jussà et al., 2022) as one of the state-of-the-art enc-dec translation models. As NLLB is trained at the sentence-level and not intended for document-level translation, we include only its sentence-level scores.

General translation assessment data. We evaluate on IWSLT2017 TED data (Cettolo et al.,

	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	–	–	–	–	–	–	–	–	–
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
NLLB 3.3B	**76.3	**64.0**	36.9	–	–	–	–	–	–	–	–	–
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 2: COMET, generative (GPRO) and contrastive (CPRO) pronoun accuracy on ContraPro, with random, structurally perturbed and gold context, and generic prompt. Random guess acc.: 33.3% EN→DE, 50% EN→FR.

2012), in English to German (EN→DE) and English to French (EN→FR). For EN→DE, we combine `tst2016-2017` for a testset of 2,271 sentences across 23 documents. For EN→FR, we use `tst2015`, containing 1,210 sentences in 12 documents.

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. We use a context size of 5 in all experiments, unless stated otherwise.

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 and randomly sample a balanced 2k subset of ContraPro and use a context size of 2.

Evaluation. We evaluate translations using BLEU (Papineni et al., 2002), CHRF (Popović, 2015), and COMET (Rei et al., 2022), 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).

2.3 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. 1.[1]

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

3 Results and Discussion

3.1 Translation Assessment

Table 1 shows the translation performance (BLEU, COMET) on IWSLT2017. CHRF, deferred to Appendix C, shows similar trends. Table 2 shows the generative and contrastive pronoun accuracy and translation performance (COMET) on ContraPro.

From Table 1, we see that **document-level *generic* prompting improves translation and pronoun resolution performance of all models over the sentence-level baseline**. Moreover, ***explicit* prompting improves instruction-finetuned mod-**

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

els' performance, while strong base-models (such as TowerBase 13B) degrade in performance.

For models under consideration in this work, translation finetuned LLMs outperform enc-dec models at overall translation. From Table 1, for EN→DE, TowerInstruct 13B outperforms all models in all context setups. For EN→FR, TowerBase 13B and TowerInstruct 13B perform comparably, with TowerBase leading in lexical metrics (CHRF,BLEU), and TowerInstruct in neural metrics (COMET).

There is room to improve LLMs' translation finetuning to better handle context-dependent discourse phenomena. As in Table 2, both enc-dec baselines outperform all LLMs in GPRO and COMET performance. Since we don't have access to NLLB and Llama-based models' training data, we cannot guarantee that ContraPro is unseen and thus that the evaluation is fair. In particular, NLLB's performance far above chance at sentence level may be due to such contamination, as sentence-level prompting forces it to *guess* the pronoun gender 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 enc-dec model in EN→DE which doesn't align with their GPRO and COMET performance on the data.

3.2 Perturbation Analysis

All models show the same degree of degradation in translation and pronoun resolution performance with a perturbed context, which is expected as no model is explicitly trained for context utilization. Looking at models' performance with total random tokens, we find that on IWSLT data, Tower models (the best at translation) are robust to random context and only degrade slightly in performance. Similarly, Tower models show the least difference in GPRO performance between gold and random contexts among all LLMs. Robustness to random context can be linked to lack of proper context utilization. We hypothesize that **explicit context-aware finetuning could improve models' handling of context-dependent discourse phenomena.**

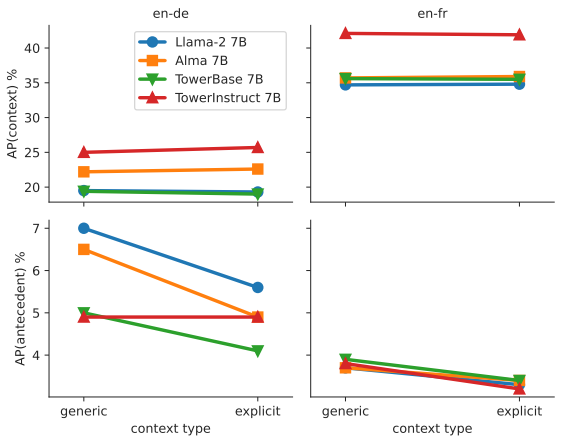

Figure 2: Attribution percentage (Eq. 1), from ALTI-Logit, assigned to the context tokens when force-decoding the correct pronoun in the ContraPro data.

3.3 Attribution Analysis

We analyze models' internals to see how much the relevant context contributes to the outputs. Figure 2 presents attribution percentages of antecedent tokens (relevant context) as well as of the whole context using ALTI-Logit. Input-erasure method, deferred to Appendix D, shows similar trends.

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

There is a need to explicitly finetune translation LLMs to focus on *relevant context.* Looking at both attribution methods, we see that for TowerInstruct 7B model (the best at translation among the 4 models tested), overall context tokens have the highest attribution percentage to the output, but when focusing on relevant parts of the context, we see that it is not the case.

4 Conclusion

We use interpretability tools (perturbation and attribution techniques) to analyze LLMs' context-utilization in document-level translation. Our experiments suggest that finetuning LLMs to translation pushes state-of-the-art translation performance beyond enc-dec 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.

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 and we leave it to future work to experiment on general translation on other language families.

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 enc-dec 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.

Ethics Statement

Nowadays, machine translation is a widely adopted technology, sometimes in sensitive, high-risk settings. Even though we propose an 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.

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.

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.

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

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.

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.

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.

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.

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.

A Training Details of the Concatenation Encoder-Decoder Model

A.1 Model

For both language pairs, 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 models are 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).

A.2 Data

For both language pairs, the models are trained on the training subset of IWSLT2017 TED data (Cettolo et al., 2012).

B Example Prompts

Fig. 4 shows examples of explicit prompts used in the perturbation experiments. We show examples of random, perturbed and gold context setups.

C CHRF Results

Table 4 shows the CHRF results in the sentence-level baseline setup, the generic prompt setup, and the explicit prompt setups.

D Input-Erasure Attribution

Figure 3 presents the attribution percentages of antecedent tokens (the relevant part of the context) as well as of the whole context using input-erasure method.

E Compute Budget

As we are only performing inference on the models and not changing any parameters, the main bottleneck is storing the model parameters. For experiments with 7B parameter models we use one A100-40GB GPUs. For 13B parameter models, we use two A100-40GB GPUs. For the attribution experiments, as ALTI-Logits works by aggregating information across layers, it is time consuming, We used around 190 GPU hours to obtain the attribution results.

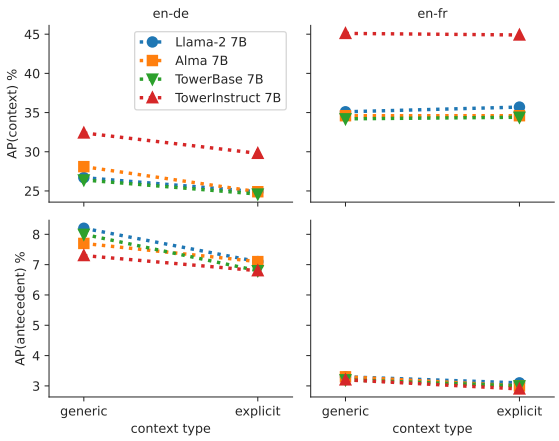

Figure 3: Attribution percentage (Eq. 1), from input-erasure, assigned to the context tokens when force-decoding the correct pronoun in the ContraPro data.

F Evaluation-Metrics Signatures

As (Post, 2018) points out the importance of providing SacreBLEU signatures for reproducability, the details of our metrics are in Table 3.

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 3: Evaluation-metrics signatures

	Sentence	Generic prompt			Explicit prompt		
	baseline	random	perturbed	gold	random	perturbed	gold
EN→DE							
Concat Enc-Dec	53.0	50.7	53.0	53.1	–	–	–
NLLB 3.3B	**59.7**	–	–	–	–	–	–
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**	–	–	–	–	–	–
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.

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 downloadingcoreżDescription 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 4: The figure shows example prompts used in the perturbation experiments, the reference translation is shown in green.