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 (randomized) document context. We find that the strongest translation LLMs are robust to random context in translation performance. However, improved document-translation performance is not always reflected in pronoun translation performance. We highlight the need for context-aware finetuning of LLMs 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 stimulated a good deal of machine translation research in the 1990s, aimed at endowing machine–translated target texts with the same document and discourse properties as their source texts (Nash-Webber et al., 2013). Since then, there has been a growing interest in document-level translation. Research efforts focused on document-level influences on lexical choice, methods and annotated resources for discourse-level MT, discourse-sensitive assessment metrics, and specific discourse phenomena in machine translation (Popescu-Belis et al., 2019).

Large language models (LLMs) show promise on multiple language tools, with recent models specially finetuned for machine translation (Alves et al., 2024; Xu et al., 2023). Wang et al. (2023) suggest translation LLMs have potential on the document level as well. While such work focuses on automatic translation metrics such as BLEU, our work investigates how those models utilize the context when performing 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.

To assess models' context utilization performance, we compare their translation performance with a random context against the gold document context. For a finer grained evaluation, we look at models' internals using attribution methods (Ferrando et al., 2023) in order to quantify the contribution of the context to the relevant translation. To the best of our knowledge, we are the first to explore context utilization in translation LLMs through perturbation and attribution methods. Our findings can be summarized as follows:

- The best translation-finetuned LLMs are robust to random context and can translate well even when prompted with random context.
- For EN→DE, translation improvements are not reflected in discourse phenomena performance, the best translation LLM performs worse than an encoder-decoder model.
- We highlight the further need for **context-aware finetuning** of LLMs to improve discourse phenomena performance.
- Adding natural language instructions to the prompts reduces the translation performance of LLMs that are not instruction-tuned.

2 Methodology

2.1 Models

We focus on LLMs fine-tuned for translation. 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 is further fine-tuned from TowerBase for translation-related tasks. We also analyze ALMA (Xu et al., 2023), which follows a two-step fine-tuning 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 fine-tuning on context use.[1] We consider the 7B and 13B versions of all models wherever feasible. As a non-LLM baseline, we include an encoder-decoder Transformer trained with concatenated context (details in Appendix E).

2.2 Datasets

We evaluate on IWSLT2017 TED data (Cettolo et al., 2012). We consider two language pairs in our experiments, namely English to German (EN→DE) and English to French (EN→FR). For EN→DE, we combine `tst2016-2017` resulting in a test set of 2271 sentences in 23 documents. For EN→FR, we use `tst2015` as the test set which contains 1210 sentences in 12 documents. We use a context size of 5 source-target pairs in our experiments.

For pronoun translation experiments we use ContraPro dataset (contrastive pronoun resolution), a subset of OpenSubtitles available for both language pairs (Müller et al., 2018; Lopes et al., 2020), consisting of examples with ambiguous pronouns, their correct translations, and automatic annotation of pronouns' antecedents (relevant context) needed for the resolution. We randomly sample a 2k subset of the data with antecedent distance of 1 or 2 sentences and use 2 source-target pairs as context.

2.3 Prompt Format

As observed by Wu et al. (2024) , the prompt format plays a significant role in LLMs' performance. A well-structured prompt can significantly boost models' performance. In our analysis, we use three prompt formats from Wu et al. (2024): a sentence-level baseline, a generic prompt, and an explicit prompt; all are demonstrated in Fig. 1.[2]

2.4 Assessing Translation Performance

We quantify performance with usual translation metrics alongside a pronoun-focused evaluation.

(a) Sentence-level prompt

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

(b) Generic prompt

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

(c) Explicit prompt

```
<src_lang>: <src context 1>
<tgt_lang>: <tgt context 1>
<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: Prompt formats used in our work.

Translation metrics. We report BLEU[3] (Papineni et al., 2002), ChrF[4] (Popović, 2015), and COMET[5] (Rei et al., 2022).

Generative pronoun accuracy (GPRO). Correctly translating ambiguous pronouns requires context. To assess the accuracy of LLMs at this job, we use the GenPro strategy on top of the ContraPro data (Post and Junczys-Dowmunt, 2023). To test the generative ability of models using GenPro, we decode a whole sentence from the model and evaluate whether the correct pronoun is included.

2.5 Analysis Overview

Like Mohammed and Niculae (2024), we follow a two-pronged approach, looking at translation and pronoun accuracy under a **perturbation analysis**, and examining the model mechanics through an **attribution analysis** via interpretability methods.

Perturbation Analysis. We compare the models' behavior when provided the actual, gold context versus when provided random tokens as context. The gold context contains the previous source-target pairs. To generate random context, we sample uniformly random tokens from the model's vocabulary, with the same size as the correct context.

Attribution Analysis For a finer-grained evaluation, we analyze how much LLMs utilize relevant parts of the context when translating ambiguous

[1] Since attribution methods require access to model internals, we exclude API-only LLMs such as ChatGPT.

[2] For TowerInstruct, we add a prefix to the prompt to indicate instruction following, as described in the model documentation: <|im_start|>user {**prompt**} <|im_start|>assistant.

[3] SacreBLEU signature (Post, 2018) nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.4.0

[4] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.0

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

	Sentence baseline			Generic prompt						Explicit prompt					
				random context			gold context			random context			gold context		
	COMET	BLEU	GPRO	COMET	BLEU	GPRO	COMET	BLEU	GPRO	COMET	BLEU	GPRO	COMET	BLEU	GPRO
EN→DE															
Concat Enc-Dec	75.4	23.4	**81.4**	68.2	20.2	48.1	75.4	23.3	**79.4**	–	–	–	–	–	–
Llama-2 7B	79.0	20.9	17.1	42.6	01.5	04.8	81.2	22.0	37.2	77.9	20.1	16.0	81.2	22.8	38.9
Llama-2 13B	76.0	02.1	13.8	56.8	06.0	07.3	82.8	25.5	37.9	78.4	22.5	19.0	76.4	01.7	22.5
TowerBase 7B	82.8	25.9	26.9	82.1	25.7	20.9	83.8	25.7	41.9	83.0	26.3	34.4	81.9	26.2	40.5
TowerBase 13B	82.7	27.1	14.0	83.5	27.3	11.5	85.0	28.8	43.7	83.4	27.2	25.5	78.3	25.8	38.5
ALMA 7B	82.9	24.8	40.2	77.1	15.7	19.7	83.4	25.3	45.2	82.4	23.4	38.9	83.7	24.5	49.5
ALMA 13B	83.8	26.2	37.6	73.7	17.3	19.9	84.3	27.1	44.9	73.7	25.6	42.9	83.4	27.1	48.8
TowerInstruct 7B	84.8	27.3	**46.9**	84.4	26.6	42.9	85.2	27.5	**50.5**	84.4	26.4	46.6	85.0	27.1	**50.7**
TowerInstruct 13B	**85.1**	**28.5**	**46.9**	84.8	27.2	39.0	**85.6**	**29.1**	45.6	84.9	27.5	45.2	**85.4**	**28.7**	46.8
EN→FR															
Concat Enc-Dec	77.8	35.8	20.2	65.8	27.9	21.6	77.6	35.6	28.5	–	–	–	–	–	–
Llama-2 7B	81.6	33.2	12.5	29.5	01.2	01.1	82.6	34.8	32.2	80.9	31.6	03.3	82.5	31.4	28.9
Llama-2 13B	77.0	17.1	06.2	54.7	04.2	01.9	84.5	38.4	33.7	81.1	34.2	03.6	83.4	06.3	15.9
TowerBase 7B	84.8	39.8	14.4	83.8	37.1	07.2	79.0	36.3	35.9	84.4	40.0	10.5	76.4	35.3	31.7
TowerBase 13B	79.5	39.6	09.4	84.9	41.0	03.1	85.9	**42.0**	36.2	85.1	40.7	07.6	69.6	31.9	35.1
ALMA 7B	80.8	28.7	13.4	52.2	07.1	03.0	81.1	27.9	28.6	80.3	28.9	06.6	81.3	30.5	27.7
ALMA 13B	83.0	33.7	14.7	60.0	10.0	04.1	83.4	33.1	31.3	82.9	33.9	09.3	83.7	35.1	31.7
TowerInstruct 7B	85.8	38.1	**34.9**	85.5	35.4	13.2	86.0	39.6	**41.2**	85.4	36.1	17.0	85.9	39.2	**39.8**
TowerInstruct 13B	**86.2**	**39.9**	**34.9**	86.0	39.3	13.3	**86.4**	41.0	39.0	86.0	39.5	16.9	**86.2**	**40.8**	38.4

Table 1: Translation performance (COMET and BLEU on the IWSLT test data, and generative pronoun accuracy (GPRO) on the ContraPro data, with or without context perturbation, for the prompts considered.

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 explanation methods often disagree. ALTI-Logit tracks the logit contributions back to the model's input by aggregating across layers and also 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 must measure how much of the overall attribution goes to a subset of the input $S \subseteq X$. This motivates the **attribution percentage**:

$$\mathrm{AP}(S)\% \coloneqq \frac{\sum_{t \in C} a_t}{\sum_{t \in X} a_t} \times 100\%. \qquad (1)$$

We use the ContraPro data and setup, force-decoding up to the pronoun, and measuring the attribution percentage of the entire context and the supporting context.

3 Results and Discussion

Table 1 shows the **translation performance** (BLEU, COMET) and the **discourse phenomena performance** (GenPro, abbreviated GPRO) of LLMs when prompted with generic and explicit context in both gold and random context setups. ChrF, deferred to Appendix C, shows similar trends. Figure 3 presents the attribution percentages of antecedent tokens (the relevant part of the context) as well as of the whole context.

Document-level prompting of LLMs improves performance compared to sentence-level prompting: Comparing the sentence-baseline results to the context-aware results, it can be seen that document-level prompts are better than sentence-level prompts in both translation performance and discourse phenomena performance. The best translation model overall is TowerInstruct 13B model, followed by TowerBase and ALMA. All the finetuned models are better than Llama-2, which is pretrained mainly on English text and thus may not be sufficient for the task; it nevertheless is competitive with the encoder-decoder baseline.

Translation finetuned LLMs are better than encoder-decoder models at overall translation, but not necessarily stronger at translating ambiguous pronouns: for EN→DE, the encoder-decoder model is much better at translating pronouns compared to all LLMs.[6] Moreover, while almost all the 13B parameter model versions are better than the 7B versions on translation metrics, this is not true on pronoun accuracy, where the best

[6]This could partly be due to data imbalance; our test subsample contains 95% examples were the target pronoun is *es*, which is the common translation for *it*.

model is TowerInstruct 7B. This suggests that there is room to improve LLMs' translation finetuning to better handle discourse phenomena.

Explicit prompting decreases translation performance for models that are not instruction tuned: Comparing the TowerBase model's performance in the two prompt formats, we see better translation and GenPro performance using the generic prompt compared to the explicit prompt. This is expected, as the model has not been trained explicitly on instruction following. In contrast, TowerInstruct is robust to the prompt format and performs comparably in both prompt formats. Llama-2 is very sensitive to the prompt format.

The Tower models are robust against random context in translation performance, but discourse phenomena performance decreases: Tower models do not exhibit substantial degradation when prompted with random context in either prompt formats. In fact, TowerBase model even shows an increase in translation performance. On the contrary, on GenPro we see a large drop in performance with random context. This reaffirms the need to evaluate on fine-grained phenomena.

Attribution percentages do not vary much across models: Unlike the larger differences in *supporting 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 models as seen in Fig. 3. The same conclusions can be drawn from input-erasure attributions (Appendix B).

Overall, our analysis shows that not only is document context necessary for marked discourse phenomena (GenPro), but it can also help improve translation performance under general metrics. Additionally, we show that the best translation-finetuned LLM (TowerInstruct) is robust to noise in the context and can produce translations that score better compared to an encoder-decoder translation model even when prompted with random context. However, focusing on pronoun translation, the situation strongly differs by language.

4 Related Work

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. Inspired by this line of research, we evaluate context utilization of LLMs as a possible new paradigm for context-aware translation.

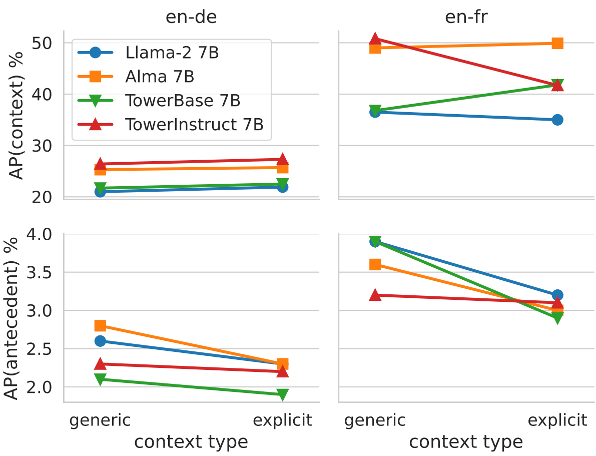

Figure 2: Attribution percentage (Eq. 1), from ALTI-Logit, assigned to the context tokens when force-decoding the correct pronoun in the ContraPro data.

Zhao et al. (2024) outline multiple interpretability techniques to analyze LLMs via mechanistic interpretability and representation engineering. Continuing the efforts on LLMs interpretability, we focus on investigating LLMs context utilization capabilities using input perturbation and attribution techniques.

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 documents (Wu et al., 2024), or a mix of sentences and documents (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.

5 Conclusion

We apply two interpretability tools (perturbation and input attribution techniques) to analyze the context-utilization ability of LLMs in document-level translation. Our experiments suggest that finetuning LLMs to translation help push the state-of-the-art translation performance beyond encoder-decoder transformer models. However, we highlight that when looking at the specifics (discourse phenomena performance), LLMs show room for improvement. We suggest 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 to being able to compute attributions of input tokens. Thus, we only used open-source LLMs in our study.

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.

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.

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.

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.

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.

Haiyan Zhao, Fan Yang, Bo Shen, Himabindu Lakkaraju, and Mengnan Du. 2024. Towards uncovering how large language model works: An explainability perspective. *Preprint*, arXiv:2402.10688.

A 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.

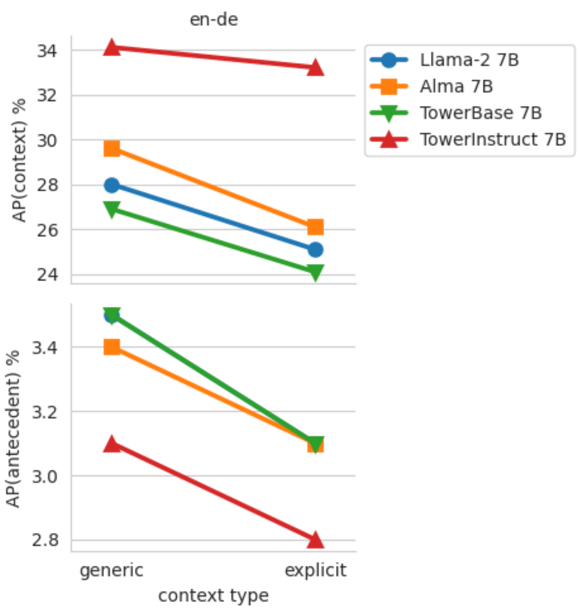

Figure 3: Attribution percentage (Eq. 1), from input-erasure, assigned to the context tokens when force-decoding the correct pronoun in the ContraPro EN→DE data.

B Erasure Attribution Percentages

C ChrF Results

Tables 2 to 4 show the ChrF results in the sentence-level baseline setup, the generic prompt setup, and the explicit prompt setups, respectively.

D Example Prompts

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

E Training Details of the Concatenation Encoder-Decoder Model

E.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).

E.2 Data

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

	ChrF
EN→DE	
LLAMA-2 7B	51.2
LLAMA-2 13B	35.0
TOWERBASE 7B	57.0
TOWERBASE 13B	57.8
ALMA 7B	54.9
ALMA 13B	56.7
TOWERINSTRUCT 7B	64.2
TOWERINSTRUCT 13B	**65.2**
EN→FR	
LLAMA-2 7B	59.2
LLAMA-2 13B	60.3
TOWERBASE 7B	65.5
TOWERBASE 13B	**64.5**
ALMA 7B	56.6
ALMA 13B	59.9
TOWERINSTRUCT 7B	64.3
TOWERINSTRUCT 13B	65.3

Table 2: ChrF scores of the sentence-level baseline on IWSLT2017 test data.

	ChrF	
setup	rand	correct
EN→DE		
LLAMA-2 7B	12.1	52.2
LLAMA-2 13B	17.9	54.8
TOWERBASE 7B	56.7	56.4
TOWERBASE 13B	57.9	59.1
ALMA 7B	46.6	54.8
ALMA 13B	43.5	56.8
TOWERINSTRUCT 7B	57.4	58.1
TOWERINSTRUCT 13B	58.2	**59.4**
EN→FR		
LLAMA-2 7B	06.5	60.1
LLAMA-2 13B	15.1	63.2
TOWERBASE 7B	64.6	59.2
TOWERBASE 13B	66.2	**66.6**
ALMA 7B	20.4	55.8
ALMA 13B	20.4	55.8
TOWERINSTRUCT 7B	63.0	65.1
TOWERINSTRUCT 13B	64.9	65.9

Table 3: ChrF scores of gold vs. random context on IWSLT2017 test data with a generic prompt.

	ChrF	
setup	rand	correct
EN→DE		
LLAMA-2 7B	51.0	53.2
LLAMA-2 13B	52.2	33.9
TOWERBASE 7B	57.1	56.5
TOWERBASE 13B	57.9	57.3
ALMA 7B	54.5	55.4
ALMA 13B	56.2	57.3
TOWERINSTRUCT 7B	57.4	57.9
TOWERINSTRUCT 13B	58.2	**59.1**
EN→FR		
LLAMA-2 7B	59.0	59.6
LLAMA-2 13B	60.0	51.7
TOWERBASE 7B	65.5	58.2
TOWERBASE 13B	66.0	55.4
ALMA 7B	56.6	57.9
ALMA 13B	59.7	61.4
TOWERINSTRUCT 7B	63.3	64.9
TOWERINSTRUCT 13B	64.9	**65.6**

Table 4: ChrF scores of gold vs. random context on IWSLT2017 test data with an explicit-context prompt.

```
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: ro practicevalue downloadingcorežDescription 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.
```

(b) Random-context prompt

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