Word Alignment as Preference for Machine Translation

Anonymous ACL submission

Abstract

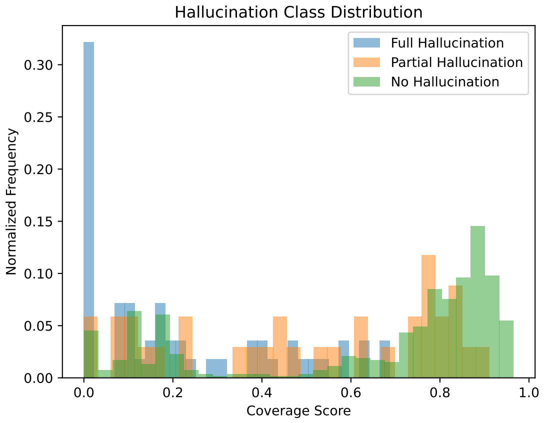
The problem of hallucination and omission, a long-standing problem in machine translation (MT), is more pronounced when a large language model (LLM) is used in MT because an LLM itself is susceptible to these phenomena. In this work, we mitigate the problem in an LLM-based MT model by guiding it to better word alignment. We first study the correlation between word alignment and the phenomena of hallucination and omission in MT. Then we propose to utilize word alignment as preference to optimize the LLM-based MT model. The preference data are constructed by selecting chosen and rejected translations from multiple MT tools. Subsequently, direct preference optimization is used to optimize the LLM-based model towards the preference signal. Given the absence of evaluators specifically designed for hallucination and omission in MT, we further propose selecting hard instances and utilizing GPT-4 to directly evaluate the performance of the models in mitigating these issues. We verify the rationality of these designed evaluation methods by experiments, followed by extensive results demonstrating the effectiveness of word alignment-based preference optimization to mitigate hallucination and omission.

1 Introduction

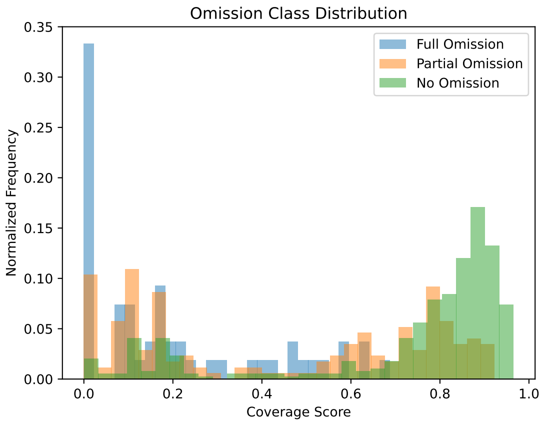
Large language models (LLMs) have been evolving rapidly and showing predominant performance in many natural language processing (NLP) tasks (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023). However, in machine translation (MT), the use of a decoder-only LLM is still limited due to issues such as model size (Xu et al., 2024a) and low-resource languages (Hendy et al., 2023). Conventional encoder-decoder MT models trained on parallel corpora still dominate in practice (Costa-jussà et al., 2022). One of the primary concerns of applying an LLM to MT is reliability. Although it does not happen frequently, an LLM is known to hallucinate (Dhuliawala et al., 2023; Zhang et al., 2023a; Bang et al., 2023) as it is pretrained to predict the next token in very large-scale raw texts. Specifically in MT, LLM-based translation systems therefore could have the phenomena of hallucination and omission, which is also a long-term challenge in the field of MT (Yang et al., 2019; Vamvas and Sennrich, 2022), known as over- and under-translation. In this work, we attempt to mitigate the hallucination and omission in LLM-based MT to improve its practicality.

Hallucination in MT occurs when information not present in the source text is generated in the translation, and omission occurs when some of the information in the source text is missed in the translation. As a related tool that explicitly aligns the source text and translation at the word level, word alignment is potentially positive for MT due to the nature of align and translate (Bahdanau et al., 2015). The degree of coverage of the source text in translation could be a direct signal to identify the hallucination and omission in MT (Tu et al., 2016). Figure 1 shows the normalized frequency of the coverage scores predicted by a word aligner. The examples that are annotated as "no hallucination or omission" tend to have a higher coverage score, while those in "full hallucination or omission" are more likely to have an extremely low coverage score. "small hallucination or omission" and "partial hallucination or omission" distribute in the middle. As the annotations are carefully made by humans and highly correlates to the coverage scores from the word aligner, this indicates that word alignment is a simple but promising direction to mitigate these phenomena.

Consequently, we propose Word Alignment Preference (**WAP**) that utilizes word alignment as a signal to optimize LLM-based MT models. WAP consists of three steps: diverse translation collection, preference data construction, and preference optimization. Specifically, we collect diverse translations with multiple existing translation tools, se-



(a) Coverage distribution of different hallucination degree.



(b) Coverage distribution of different omission degree.

Figure 1: A preliminary experiment shows that higher coverage scores correlates to less hallucination and omission. The coverage scores are predicted by a word aligner (Wu et al., 2023). The human annotation of hallucination and omission is from HalOmi benchmark (Dale et al., 2023b). Details about the dataset and word alignment model can be found in §3.1.

lect chosen and rejected examples with the word aligner (Wu et al., 2023), and optimize the model on preference data using direct preference optimization (DPO) (Rafailov et al., 2024).

Furthermore, the evaluation of hallucination and omission is challenging, and there is no existing evaluator specifically designed for this. Improving the BLEU and COMET score does not necessarily mean reducing hallucination and omission because there are other factors such as mistranslation and fluency. In addition, hallucination is relatively infrequent, although very severe once happens. Hence, to effectively evaluate it, we design extensive experiments that include testing on instances that potentially have the problem of hallucination and omission, and using GPT-4 as the evaluator with comprehensive analysis. Experimental analysis demonstrates the effectiveness of WAP in mitigating hallucination and omission in MT.

In summary, the contributions of this work include the following:

- We studied the correlation between the coverage score by word alignment and the phenomena of hallucination and omission in MT. From the preliminary experiments in Figure 1 we found that word alignment is a promising signal to mitigate it.
- In §2 we propose a novel approach, namely WAP, to construct a word alignment-based preference dataset, and use DPO to optimize the LLM-based MT model. The validity of the preference dataset is also demonstrated by direct fine-tuning on preferred and rejected translations in §4.4.
- As there is no benchmark particularly for evaluating the performance of MT models on hallucination and omission. We design various experiments, including selecting hard instances and utilizing GPT-4 as the evaluator in §3.2. The effectiveness of the evaluation, as well as the proposed WAP has been validated through experiments and analysis in §4

2 Proposed approach

2.1 Gathering translation candidates

To steer the MT model to avoid hallucination and omission using preference optimization, we first need comparable but different translations. Starting with a source text x, we utilize K methods to produce translations, notated as $\pi^1, ..., \pi^K$. Then we can get a set of translations Y, in which $y^k \in Y$ is obtained by $y^k = \pi^k(x)$ and $|Y| = K$.

Details of gathered translations We start with the parallel training data in ALMA (Xu et al., 2024a). This parallel data encompasses five language pairs with human translations in both directions: $cs \leftrightarrow en$, $de \leftrightarrow en$, $is \leftrightarrow en$, $zh \leftrightarrow en$ and $ru \leftrightarrow en$. We employ ISO 639 language codes[1] to denote languages. Specifically, "cs" corresponds to Czech, "de" to German, "is" to Icelandic, "zh" to Chinese and "ru" and "en" to Russian and English, respectively. To generate the translations we require, this dataset is translated in both directions

[1] https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

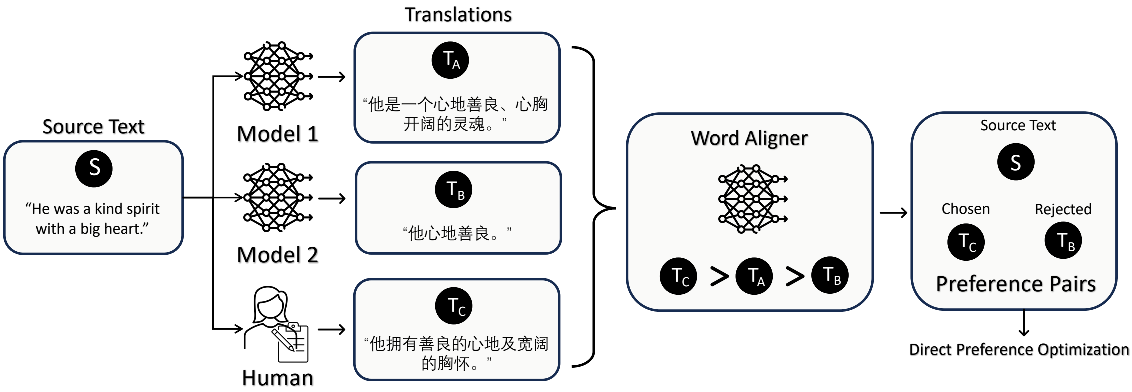


Figure 2: An illustration of WAP framework. The source is first translated by multiple MT tools, including human translation. An external word aligner is then utilized to predict the coverage score for each translation. Finally, translation with the highest and lowest coverage score are selected as preference pairs for preference optimization.

using two well-known MT tools, including DeepL[2] and ChatGPT (`gpt-3.5-turbo-0613`)[3]. The prompt for ChatGPT that we utilize to translate sentences is shown in Figure 6. The original human-written translation in the training set is also utilized. In particular, Icelandic (*is*) is not supported by the DeepL API, therefore, we use the Google Translate API[4] as an alternative.

2.2 Selecting chosen and rejected translation

After obtaining the translation candidates $(y^1, ..., y^K)$, we use a state-of-the-art public word aligner, namely WSPAlign[5], to automatically annotate the degree of coverage for each translation. We follow the usage setting in the original paper of WSPAlign (Wu et al., 2023). In particular, WSPAlign performs a bidirectional alignment and uses a threshold to filter out low-confident alignment of word pairs. Then, the ratio of the source words, *that are aligned with at least one word*, in the translation is taken as the coverage score, which will be used for the following preference annotation. The whole process predicting the coverage score is notated as $\mathrm{C}(\cdot, \cdot)$. Formally, the coverage score for a translation y^k can be calculated by $\mathrm{C}(x, y^k) \in [0.0, 100.0]$. Subsequently, the preferred translation and the rejected translation are selected as follows:

$$
\begin{aligned}
y^w &= \arg\max_{y^k \in Y} \mathrm{C}(x, y^k) \\
y^l &= \arg\min_{y^k \in Y} \mathrm{C}(x, y^k)
\end{aligned}
\tag{1}
$$

where y^w is the chosen translation and y^l is the rejected one. Then a triplet (x, y^w, y^l) is constructed for the following preference optimization.

2.3 Filtering

Note that the whole pipeline of constructing the preference data is automatic, and existing MT and word alignment models are not perfect. Even for human-annotated translation, the quality of it is also an issue that cannot be ignored (Xu et al., 2024b), and may affect the performance of the model trained on it. Hence, noises are inevitable in both the translated texts and the preference choices. On the other hand, the MT tools we choose generally have good performance, it could happen that the generated translations are not diverse enough, leading to the preference signal being disrupted. To improve the quality of the constructed preference datasets as much as possible, multiple strategies are applied to filter out potential bad training instances:

- Remove the instance when the chosen and rejected translations only have a marginal difference in coverage score. The difference threshold is empirically set as 5.0, that is, (x, y^w, y^l) is excluded from the dataset if $\mathrm{C}(x, y^w) - \mathrm{C}(x, y^l) < 5.0$.

- Remove the instance when the chosen and rejected translations are too semantically similar.

[2]`https://www.deepl.com/en/translator`
[3]`https://platform.openai.com/docs/models/gpt-3-5-turbo`
[4]`https://cloud.google.com/translate/docs/basic/translate-text-basic`
[5]`https://github.com/qiyuw/WSPAlign`

Sentence embedding is a widely used technique for sentence similarity with low computation cost (Gao et al., 2021; Wu et al., 2022; Zhao et al., 2024). LaBSE (Feng et al., 2022)[6] is used in our experiments. We notate it as $\mathrm{LB}(\cdot)$. The similarity threshold is empirically set as 0.9, i.e. (x, y^w, y^l) is excluded from the dataset if $\mathrm{sim}(\mathrm{LB}(y^w), \mathrm{LB}(y^w)) > 0.9$. $\mathrm{sim}(\cdot, \cdot) \in [0.0, 1.0]$ is cosine similarity.

- One possible failure case for word alignment is when the MT models directly copy the original texts, which is bad translation, but gets a high alignment score because the wrong translation is partially the same with the original texts. To remove this part of the noise, we calculate the BLEU score (Papineni et al., 2002)[7] for the chosen translation and exclude it if the BLEU score > 20.0.

The details and analysis of the fianl preference dataset after filtering is introduced in §3.1.

2.4 Optimization LLM-based MT model

The final step is to optimize the LLM-based MT model on our preference data. Direct preference optimization (DPO) (Rafailov et al., 2024) is a simple but effective approach that directly optimizes the preference model on a pre-constructed static dataset. DPO has been applied to optimize LLM in preference data (Tunstall et al., 2023; Xu et al., 2024b) recently. We also utilize DPO as an optimization approach. Formally, the training objective is as follows,

$$l = -\log \sigma(\beta \log \frac{\pi(y^w|x)}{\pi_{ref}(y^w|x)} - \beta \log \frac{\pi(y^l|x)}{\pi_{ref}(y^l|x)}) \tag{2}$$

where σ is the sigmoid function, π is the model to optimize and π_{ref} is the reference model. We use ALMA-13B[8] as our base model, i.e., the starting point of π, in the experiments. ALMA-13B is also used as a reference model π_{ref}, but note that π_{ref} will not be updated during training.

3 Evaluation

The experimental setup is introduced in §A.

[6]https://huggingface.co/sentence-transformers/LaBSE

[7]https://github.com/mjpost/sacrebleu

[8]https://github.com/fe1ixxu/ALMA

3.1 Baselines and evaluation datasets

We choose ALMA-13B[9] as the baseline for all experiments in this paper, as well as the starting point of optimization. ALMA (Xu et al., 2024a) was trained from Llama (Touvron et al., 2023) in two steps: initial fine-tuning on monolingual data and subsequent fine-tuning on a small set of high-quality parallel data.

For fairly studying the effect of word alignment preference, we use the data used in the supervised fine-tuning in ALMA as the source dataset to construct our preference data in §2. Specifically, the source data was collected from WMT'17 (Bojar et al., 2017) to WMT'20 (Barrault et al., 2020), in addition to the development and text dataset from Flores-200 (Costa-jussà et al., 2022). After filtering, we finally make 20,074 and 2,226 preference triplets for training and development, respectively. For evaluation, the test set is from WMT22, except that $is \leftrightarrow en$ is from WMT21. The remaining data from WMT21 (except $is \leftrightarrow en$) is used as the development set. Specifically, 3485, 4021, 2000, 3912, 4053 examples are included in the test set for $cs \leftrightarrow en$, $de \leftrightarrow en$, $is \leftrightarrow en$, $zh \leftrightarrow en$, and $ru \leftrightarrow en$, respectively.

HalOmi In particular, we want to validate whether our proposed method is capable of mitigating hallucination and omission in MT. Hence, we also utilize HalOmi (Dale et al., 2023b) in the experiments. HalOmi is an evaluation benchmark for the detection of hallucination and omission in MT. It contains fine-grained sentence-level and token-level annotations of full and partial hallucinations and omissions that cover 18 language directions. Each instance in the data set was annotated in "No hallucination and omission", "Small hallucination and omission", "Partial hallucination and omission" or "Full hallucination and omission" by humans. In this paper, we use it to test the performance of GPT-4 as an evaluator. Details are in §3.2.

3.2 The design of evaluation

We focus on optimizing LLM-based MT models to avoid hallucination and omission. However, to our best knowledge, there is no benchmark measuring MT models specifically for this issue, making the evaluation very challenging. Improving the BLEU or COMET score does not necessarily mean reducing hallucination and omission because there are

[9]https://huggingface.co/haoranxu/ALMA-13B

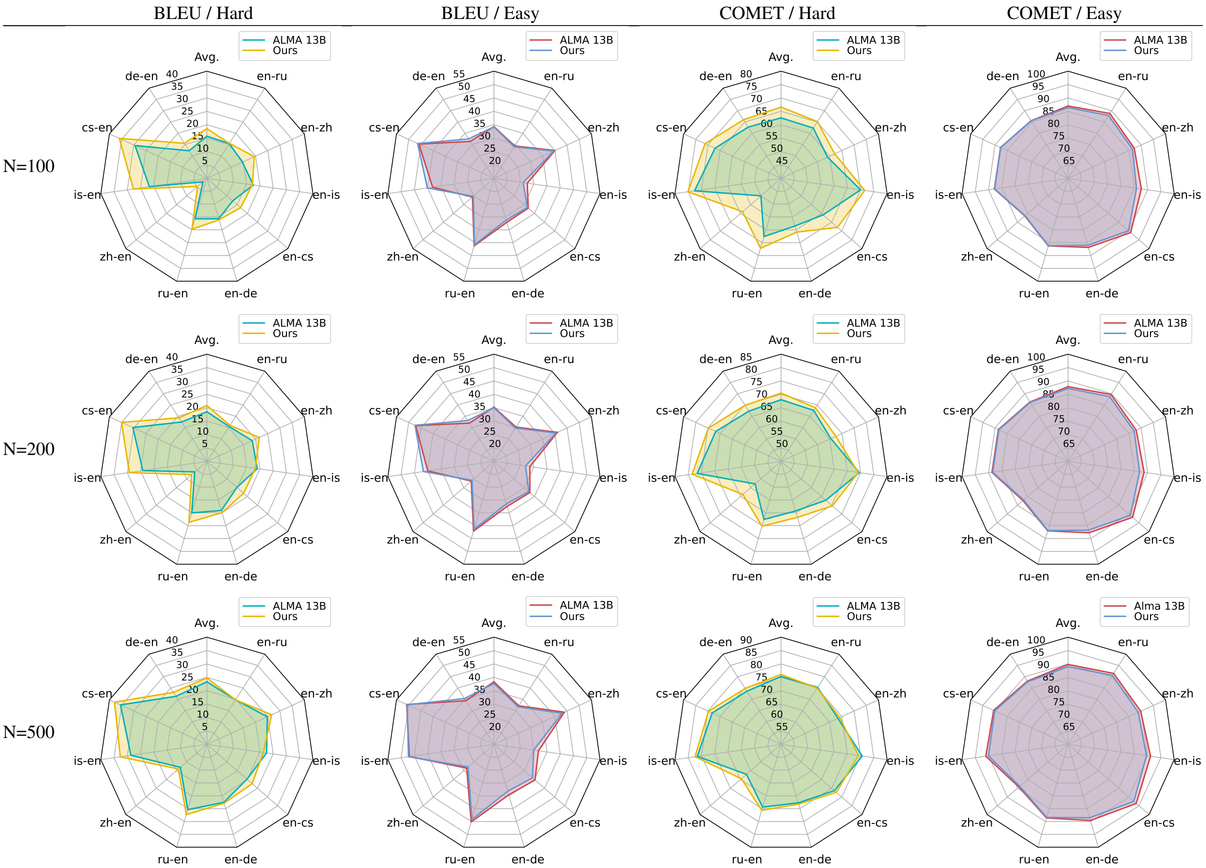


Figure 3: Comparison of WAP and baseline in hard and easy instances. N instances with the lowest COMET score by the baseline are selected from the test set as hard instances, and the remaining are easy instances. Results when $N = 100$, 200 and 500 are presented. Refer to §D for the full numeric results of the entire test.

other factors such as mistranslation and fluency. In addition, hallucination is relatively infrequent, although very severe once happens. To intuitively validate whether our approach is capable of mitigating hallucination and omission in MT, we design several evaluation strategies in this section.

Select hard instances. We first select instances that the baseline model does not perform well on. This subset of instances is labeled as *hard instances* in this paper. The subset of the remaining examples is labeled as *easy instances*. Specifically, N instances with the lowest COMET score are selected from the test set for each translation direction. As hard examples tend to include more hallucination and omission, we report the comparison of models on hard examples and remaining examples, respectively. In the experiment, we sample three subsets where $N = 100$, $N = 200$ and $N = 500$. The experimental analysis can be found in §4.1. Note that the hard instances are only selected for evaluation. We do not differentiate hard or easy instances in the training set. Only word alignment signal is used to select preferred dataset for a fair comparison.

	Hallucination			Omission		
	No	Partial	Full	No	Partial	Full
# of examples	817	42	65	627	237	60
Avg. score	84.19	45.95	3.84	87.97	66.28	1.66
Pearson Corr.	0.5969			0.5686		

Table 1: Average coverage score calculated by GPT-4 for different level of hallucination or omission. The Pearson Correlation between the annotated labels and GPT-4 coverage scores is also reported. Ideally, higher score should correlate to less hallucination and omission.

Utilize LLM as the evaluator for hallucination and omission. Besides the BLEU and COMET in hard instances, a direct estimate of the degree of hallucination and omission in translation is still needed. As we mentioned earlier that improving the BLEU and COMET score does not necessarily mean reducing hallucination and omission because there are other factors such as mistranslation and fluency, we utilize the generalization and reasoning ability of LLM (Kojima et al., 2022; Mitchell et al., 2023; Wei et al., 2023) to achieve this direct evaluation. We use one of the most powerful LLM, `gpt-4-0613`[10], as the evaluator. LLM is prompted to check whether the given translation has hallucination or omission referring to the given source texts. A coverage score between 0 and 100 is output as the degree metric. The prompt used is shown in Figure 7.

Is LLM really capable of evaluating hallucination and omission in MT? Despite the fact that LLMs have shown impressive zero-shot performance in various tasks (Kojima et al., 2022; Mitchell et al., 2023; Wei et al., 2023), the assessment of LLM in the evaluation of hallucination and omission is still important because it has not been widely used on this task. We use HalOmi datasets introduced in §3.1 to assess this ability of GPT-4. The examples in $de \leftrightarrow en$, $zh \leftrightarrow en$, and $ru \leftrightarrow en$ are selected, then GPT-4 is used to predict the coverage score for these examples.

Table 1 shows the average score of the degree of coverage predicted by GPT-4. The examples from HalOmi are split into three subsets based on the labels. We merged the "Partial hallucination and omission" and "Small hallucination and omission" in the original because the number of examples in these two categories is small. It clearly demonstrates that examples annotated as "No hallucination and omission" have a higher coverage score predicted by GPT-4 and those in "Full hallucination and omission" have an extremely low coverage score. As a result, using GPT-4 is an effective way to assess whether a translation has the problem of hallucination or omission.

4 Experimental results

4.1 Evaluation on hard instances

In §3.2 we introduce how to select hard instances from the test set and explain why hard instances are suitable to assess hallucination and omission. In this section, we evaluate our model on these hard instances and the remaining examples, respectively. Figure 3 demonstrates the results when the number of hard instances $N = 100, 200$, and 500, respectively. The following findings can be concluded:

- WAP consistently outperforms the baseline in hard instances in most translation directions, for both BLEU and COMET metrics.
- WAP generally reaches competitive performance compared to the baseline for both BLEU and COMET.
- With increasing the number of hard instances, the improvement gained by WAP gets smaller.

These results indicate that WAP mitigates hallucination and omission to a certain extent, because these issues are more likely to occur in hard instances. In addition, with the improvement in the hard instances, our model remains competitive to the baseline in the remaining easy instances. It is reasonable that there is no significant difference in the easy instances because the compared models are generally good. The challenging part should be in the hard ones. Moreover, it is also observed that with increasing N, the improvement gets narrower. The reason is that more relatively easy instances are included in the subset. This is another evidence that WAP provides gains particularly for hallucination and omission in MT. The specific numeric results and the overall results for the entire test set are shown in §D.

4.2 Direct evaluation of hallucination and omission by GPT-4

In addition to improving hard examples, which is more likely to have hallucination and omission, direct evaluations of them are also needed to confirm the effectiveness of the proposed WAP. In §3.2 we have verified the usefulness of GPT-4 as an evaluator with experiments. In this section, we prompt GPT-4 to directly predict a coverage score as the metric of hallucination and omission. The results are demonstrated in Table 2. The reported number is the average of the coverage scores in hard examples. The results show that our model outperforms the baseline in all translation directions except $en \leftrightarrow is$. Specifically in the average score of all translation directions, WAP outperforms the baseline model by **4.96**, **1.63** and **1.24** when N=100, 200 and 500, respectively. The trend is similar to that of §4.1, which directly indicates that the LLM-based MT model is steered to avoid generating hallucination and omission in MT with the preference dataset we constructed.

4.3 Human evaluation

Although the validity of GPT-4 as evaluator for hallucination and omission has been demonstrated in §3.2 and Table 1, we conduct a human evaluation to further verify our findings, as LLM could

[10] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

	de-en	cs-en	is-en	zh-en	ru-en	en-de	en-cs	en-is	en-zh	en-ru	Avg.
N=100											
Baseline	94.30	92.95	94.90	63.08	89.85	92.85	82.75	**97.05**	84.65	90.53	88.29
+WAP	**95.85**	**94.65**	**96.05**	**80.23**	**91.75**	**96.25**	**91.85**	96.10	**92.90**	**96.87**	**93.25**(+4.96)
N=200											
Baseline	95.71	95.05	95.45	74.83	92.83	94.20	89.95	**97.70**	89.19	94.25	91.92
+WAP	**97.10**	**96.55**	**97.48**	**85.63**	**95.53**	**95.18**	**91.84**	96.73	**92.81**	**96.66**	**94.55**(+2.63)
N=500											
Baseline	97.18	96.74	97.29	87.85	96.16	97.35	94.46	98.21	91.64	96.10	95.30
+WAP	**98.10**	**97.79**	**98.12**	**90.76**	**97.82**	**97.36**	**96.05**	**98.22**	**94.07**	**97.13**	**96.54**(+1.24)

Table 2: Coverage score output by GPT-4. The range of the score is $[0.0, 100.0]$. The average score is reported for each translation direction. Higher scores are highlighted in bold.

	Translation Quality	Hallucination				Omission			
		No	Small	Partial	Full	No	Small	Partial	Full
Baseline	11.33%	64.00%	21.00%	11.33%	3.66%	56.00%	25.33%	13.66%	4.33%
+WAP	**39.66%**	**75.66%**	**17.33%**	**7.00%**	**0.00%**	**80.00%**	**16.66%**	**5.33%**	**0.00%**

Table 3: Human evaluation on "zh-en" when N=100. Translation quality is the measured by ratio of examples where WAP beats the baseline. The remaining columns present the ratio of examples in which the corresponding degree of hallucination or omission occurs. Better model is highlighted with bold fonts.

still be unreliable. The subset of "N=100" on "zh-en" is selected. Three volunteers who speak Chinese and English are asked to assess the quality of the translation and the degree of hallucination and omission for the baseline and our model, without knowing which model generates the translations. Table 3 demonstrates the results. In general, our model generates better translation in 39.66% of the examples, while the percentage for ALMA is 11.33%. Furthermore, it is observed that with DPO on word-alignment preferred data fine-tuning, the degree of both hallucination and omission decreases. Specifically, the percentage of "no hallucination" increases from 64% to 75.66%, and that of "small, partial, and full hallucination" decreases accordingly. The decrease in omission is more distinct, in which the percentage of "no omission" increase by 24%. Notably, for both hallucination and omission, the percentage of "full hallucination and omission" has decreased to 0 for our model. These results indicate that omission is more frequent than hallucination, and WAP can mitigate them in LLM-based MT model.

4.4 Ablation study

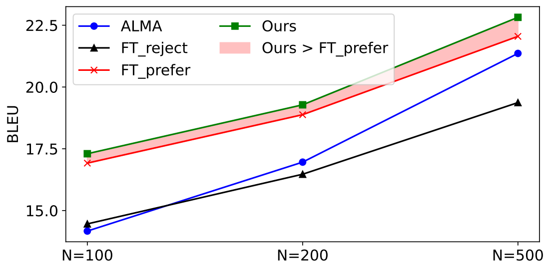
In this section, we conduct in-depth investigation for our word alignment preference, as we use the same training data as our baseline ALMA, i.e., human translation, but extra translations from DeepL and ChatGPT are included to conduct our preference data. To investigate where the improvement comes from, we introduce two variants without preference tuning to compare with WAP.

- *FT_reject*: directly fine-tuning ALMA with the rejected translations in the dataset.
- *FT_prefer*: directly fine-tuning ALMA with the preferred translations in the dataset.

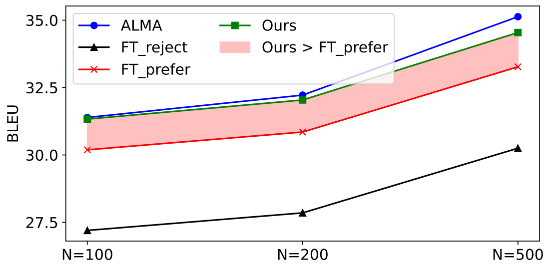
The comparison is demonstrated in Figure 4.

Does the preferred data really better contribute to the training? It is observed that *FT_prefer* significantly outperforms *FT_reject* in both hard and easy instances. This indicates that our proposed pipeline ensures that the samples are selected, leading to better translation quality.

Is the DPO preference tuning necessary? Particularly, the filled area demonstrates the necessity of preference tuning using DPO. In hard instances *FT_prefer* can reach a competitive performance with a small gap. However, in easy instances, *FT_prefer* largely underperforms WAP and ALMA, which limits the practicality of it. The possible reason for the different performance in the hard and easy instances is the direct fine-tuning. Directly fine-tuning on the preferred data without the comparison with rejected examples could cause a hard fitting to the word-aligned preference but ignore the general translation quality.



(a) Hard instances



(b) Easy instances

Figure 4: Ablation study. Results in BLEU is demonstrated. Higher BLEU is better. For fair comparison the range of y-axis are the same for hard instances and easy instances. The result in COMET is in the same trend, which can be found in Figure 8 in the Appendix.

5 Related work

Hallucination and omission in MT. Hallucinations are cases in which the model generates output that is partially or completely unrelated to the source sentence, while omissions are translations that do not include some of the input information (Dale et al., 2023b). Dale et al. (2023a) explore methods that leverage the internal workings of models and external tools, such as cross-lingual sentence similarity and natural language inference models, to detect and mitigate hallucinations in MT. HalOmi (Dale et al., 2023b) introduces an annotated dataset specifically designed to detect hallucinations and omissions. In Figure 1 and §3.2 we use HalOmi as a reference to assess how these two phenomena correlate to the coverage output of the GPT-4 evaluator and the word aligner, respectively. In particular, Yang et al. (2019) introduce using word alignment to reduce omission in MT, which partially inspires our idea.

Preference tuning for LLMs. LLMs are capable of completing tasks in the zero-shot or few-shot manner (Radford et al., 2019; Brown et al., 2020). In addition, performance in downstream tasks can also be enhanced by fine-tuning them with instruction datasets (Wei et al., 2022; Chung et al., 2024; Ouyang et al., 2022). However, acquiring instruction datasets is costly, while obtaining preferences for LLM responses is relatively easier (Rafailov et al., 2024). DPO (Rafailov et al., 2024) directly optimize LLM with preference data by removing an extra reward model. We utilize DPO in this work due to the ease of use and effectiveness. A contemporaneous preference-based MT model ALMA-R (Xu et al., 2024b), introduces contrastive preference optimization to fine-tune LLMs specifically using reference-free MT metrics and human annotation as preference. ALMA-R focuses on improving general LLM-based MT but we attempt to mitigate the hallucination and omission in MT. In addition, our preference data are entirely made automatically, which also draws the difference between ALMA-R and our work.

Word alignment. Word-level information is useful in many NLP tasks such as language pretraining (Chi et al., 2021; Wu et al., 2021), cross-lingual sentence embedding (Zhang et al., 2023b; Li et al., 2023; Miao et al., 2024), and particularly for MT (Bahdanau et al., 2015; Tu et al., 2016). Word aligners based on pre-trained language models (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Nagata et al., 2020; Chousa et al., 2020) have outperformed previous ones based on statistical MT (Och and Ney, 2003; Dyer et al., 2013). WSPAlign (Wu et al., 2023) is a pre-trained word aligner outperforming most previous ones, hence we use it in the experiments.

6 Conclusion

The problem of hallucination and omission, a long-standing problem in MT, could become more severe when an LLM is used because an LLM itself could hallucinate or omit in nature. In this paper, our aim is to mitigate this problem in LLM-based MT by optimizing the model toward a preference for better word alignment. We construct preference datasets by collecting translations using multiple MT tools and selecting the preference pair with a higher coverage score output by a word aligner. DPO is then utilized to optimize the model towards the word-aligned preference. As evaluation of hallucination and omission is challenging, we design experiments that include selecting hard instances and using GPT-4 to directly predict coverage score, ensuring an effective evaluation, which indicates that the proposed WAP mitigates hallucination and omission, especially in hard instances.

Limitation

The primary limitation of our method stems from the imperfections of the word alignment model. Within our approach, it is inevitable to encounter some alignment errors, which we address through a filtering method. However, this solution adds complexity and clutter to the method. Additionally, the effectiveness of our method is diminished for low-resource language translations due to the limited number of parallel sentences available. Lastly, our reliance on the GPT-4 API to evaluate the results introduces a significant cost factor. We aim to find a cost-free alternative for this evaluation process in future work.

Ethical Statement

All datasets and checkpoints used in this paper are copyright-free for research purposes. Previous studies are properly cited and discussed. This research aims to improve LLM-based machine translation models with word alignment preference data, and the preference is made by an automatic word aligner. We do not introduce additional bias to particular communities. We have obtained the consent of the annotation volunteers for this study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv*, abs/2309.11495.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3466–3478, Toronto, Canada. Association for Computational Linguistics.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. *arXiv preprint arXiv:2404.02490*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv*, abs/2302.10205.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. 2021. Taking notes on the fly helps language pre-training. In *International Conference on Learning Representations*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023a. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023b. Veco 2.0: Cross-lingual language model pre-training with multi-granularity contrastive learning. *arXiv preprint arXiv:2304.08205*.

Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. 2024. Leveraging multi-lingual positive instances in contrastive learning to improve sentence embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 976–991, St. Julian's, Malta. Association for Computational Linguistics.

A Experimental setup

The implementation from alignment-handbook[11] is used for the training of DPO. The learning rate is searched based on performance on development set and set to 5e-6. LoRA (Hu et al., 2021) is used. r is set as 16 and β is set as 0.1. We train the model for 1 epoch and fix the random seed to 42. The model is trained on 4 $\times$ Nvidia A100 80G and the total batch size is 64. For evaluation, we use the implementation of ALMA[12] to calculate the BLEU and COMET scores.

B Details of dataset

Figure 5 presents the varying proportions of "chosen" and "rejected" preference pairs from three sources: ChatGPT, DeepL, and Human. The figure indicates that the majority of the "chosen" translations originate from ChatGPT, while a significant portion of human-written translations are "rejected". This observation supports the conclusion that human-written translations can also exhibit quality issues, as discussed in ALMA-R (Xu et al., 2024b). Examples in our constructed preference dataset are presented in §C.1.

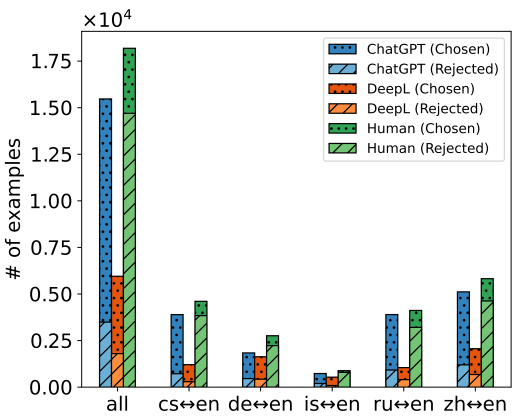


Figure 5: This figure illustrates the proportions of "chosen" and "rejected" preference pairs derived from three sources: ChatGPT, DeepL and Human. "all" represents the overall proportion for the aggregated dataset. $xx \leftrightarrow en$ is the subset pair of English and another language. Particularly, Google Translate is used for $is \leftrightarrow en$ as an alternative to DeepL.

Prompt for Translation

```
You are a helpful assistant that translates
{SOURCE_LANG} sentences to {TARGET_LANG}
sentences.

{TEXT}
```

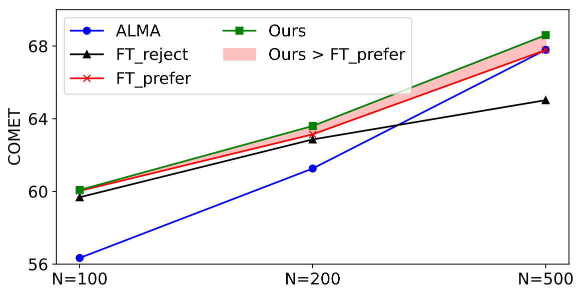
Figure 6: The prompt of ChatGPT that we use to translate sentences.

Prompt of Coverage Calculation

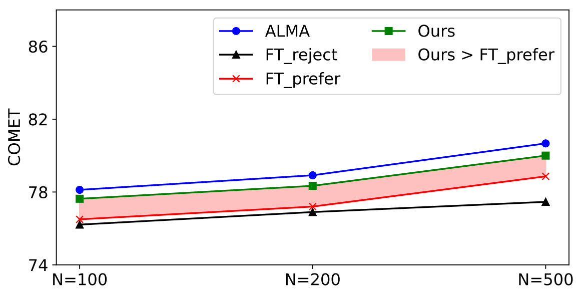
```
Given a source sentence {SOURCE_TEXT} in {
SOURCE_LANGUAGE} and a translation
{TRANSLATION} in {TARGET_LANGUAGE}, does
the translation has hallucination or
omission to the source contents?

***You MUST answer with only a coverage
percentage score***.
```

Figure 7: Prompt to calculate the coverage score.



(a) Hard instances



(b) Easy instances

Figure 8: Ablation study. Results in COMET is demonstrated. Higher COMET is better. For fair comparison the range of y-axis are the same for hard instances and easy instances. Refer to §4.4 for discussion.

Example 1 (Chinese-English)		**Coverage Score**
source	“我想，在考虑重播时，可以解决这个问题”，Coker 说道。	–
chosen (gpt-3.5)	"I think, when considering replay, this issue can be resolved," Coker said.	94.03
rejected (human)	"<<<I think that when I think about>>> the replay, <<<I think that>>> we can probably work it out," Coker said.	79.87
Example 2 (Chinese-English)		**Coverage Score**
source	<<<富勒>>>在政变图谋失败后	–
chosen (deepl)	<<<Fuller>>> after the failed coup attempt	83.76
rejected (human)	After the failure of the attempted coup,	59.59
Example 3 (English-Chinese)		**Coverage Score**
source	<<<Originally a one-bedroom property with a convoluted layout - you had to walk through the kitchen to get to the bedroom>>> - Joanne wanted to add storage space and a mezzanine to make the most of the generous ceiling height.’	–
chosen (gpt-3.5)	<<<最初是一个一居室的房产，布局错综复杂- 你必须穿过厨房才能到达卧室>>> - 然而乔安妮想要增加存储空间和一个夹层，以充分利用宽敞的天花板高度。	83.76
rejected (deepl)	乔安妮希望增加储藏空间和一个夹层，充分利用宽敞的天花板高度。	69.97

Table 4: Examples in the preference dataset. The hallucination in rejected examples and omission in the source sentence are highlighted with <<< >>>. The corresponding contents that are omitted in the rejected example are highlighted with <<< >>> in the chosen example. The coverage is calculated by word aligner, refer to §2 for details.

C Example analysis

C.1 Examples of the preference dataset

Table 4 includes three examples in our dataset, in which the source sentence, the chosen and rejected translations are shown. Refer to §B for a detailed construction of the dataset. **Example 1**: the rejected translation is from human annotation, in which it repeats the term of “I think” unnaturally. The possible reason could be the resource of the parallel data, e.g., direct collection from transcriptions. **Example 2**: “Fuller” is omitted by human annotation while translated by DeepL. **Example 3**: the chosen translation is from gpt-3.5-turbo that completely translates the source sentence. In contrast, the translation by DeepL omits the first half.

C.2 Translation examples

Table 5 shows illustrative comparison between translations from the baseline and our model. **Example 1**: “in HBO’s ’The Gilded Age’" in the source sentence is omitted by the baseline. In contrast, our model successfully translate the corresponding part into Chinese. **Example 2**: the baseline generates “卡扣 (fastening)” infinitely in translation. This type of hallucination also occurs in other LLM applications, which emphasizes the need to address the hallucination issue in LLM-based MT models. **Example 3**: “等到什么时候 (when to wait)” is omitted by the baseline model while our model translate that into “how long I have to wait” properly.

[11]https://github.com/huggingface/alignment-handbook

[12]https://github.com/fe1ixxu/ALMA

D Specific results

Table 6 shows the numeric results in Figure 3, in which boxes on a blue background highlight the cases where our model outperforms the baseline by a margin > 1.0, and the boxes in red are the opposite. Boxes without background indicate the cases when our model and the baseline have competitive performance where the margin < 1.0.

In addition to the main findings in §4.1 that our model generally performs better in harder instances, from the results it can also be observed that our model particularly performs worse on “*en-is*” than in other translation directions. The reason could be

Example 1 (English-Chinese)		Coverage Score
Source	Sunday Best: Enter 1880s New York <<<in HBO’s "The Gilded Age">>>	–
Translation (Baseline)	周日最佳：进入1880 年代的纽约	70.0
Translation (Ours)	周日最佳：进入1880 年代的纽约<<<，在HBO 的《金碧辉煌时代》>>>	100.0
Example 2 (English-Chinese)		**Coverage Score**
Source	Liner Fastening and Hanging Tabs Inner tabs are provided to keep a loose liner in position, corresponding in position with the tabs we provide on our liners.	–
Translation (Baseline)	粘贴和悬挂<<<卡扣的内部卡扣用于保持卡扣>>>	0.0
Translation (Ours)	内固定和悬挂标签内固定和悬挂标签用于保持薄膜在位，与我们提供的标签对应。	60.0
Example 3 (Chinese-English)		**Coverage Score**
Source	不知道要<<<等到什么时候>>>	–
Translation (Baseline)	I don’t know when	90.0
Translation (Ours)	I don’t know <<<how long I have to wait>>>	100.0

Table 5: Translation Examples. The hallucination in translation by the baseline and the omission in the source sentence are highlighted with <<< >>>. The corresponding contents that are omitted from the baseline are highlighted with <<< >>> in our translation. The coverage is calculated by GPT-4, refer to §3.2 for details.

that Icelandic is a low-resource language and we used external tools such as WSPAlign and Google Translate to build the training data. Hence, the relatively unreliable performance of external tools on low-resource languages can induce noises in our training data. This could be a future direction for building more reliable word alignment signals and particular research on low-resource languages.

In addition, Table 6 reports the overall performance when we do not split the dataset into the hard and easy subset. The results show that our model and ALMA have generally competitive performance. Specifically, if we only consider the margin larger than 1.0, our model outperforms ALMA on *de-en* and *is-en* in BLEU while ALMA performs better on *en-is* in both BLEU and COMET. In particular, a significance test is conducted to investigate numeric degradation when all instances are included. We utilize bootstrap sampling from example-wise COMET scores with 100,000 iterations and calculate the p-value. Based on the results of the significance test, there is no statistical significance when the margin is greater than 0.25, indicated by a p-value larger than 0.05. This suggests that our approach does not degrade the general performance by a margin of 0.25 or more, while improving that on hard instances by a large margin of 3.47. Note that the focus of this work is the problem of hallucination and omission, general metrics for MT are only partially related to our evaluation. The evaluation by LLM and humans is also important, as we discussed in §3.2.

Model-Metric	de-en	cs-en	is-en	zh-en	ru-en	en-de	en-cs	en-is	en-zh	en-ru	Avg.
						N=100					
						Easy instances					
ALMA-BLEU	31.38	45.79	38.14	25.64	41.25	32.09	31.95	27.57	40.05	29.37	31.39
Ours-BLEU	32.50	46.32	40.13	25.23	40.80	31.22	31.55	26.00	39.55	29.01	31.33
ALMA-COMET	85.57	87.71	87.82	81.38	86.26	86.84	90.90	87.61	87.14	88.80	78.12
Ours-COMET	85.50	87.67	87.71	81.24	86.17	86.02	89.84	85.80	86.39	87.89	77.63
						Hard instances					
ALMA-BLEU	12.25	29.49	21.72	1.95	15.73	15.71	12.79	17.51	14.59	15.45	14.17
Ours-BLEU	15.56	35.93	27.72	4.62	19.77	16.15	16.67	17.13	19.49	15.54	17.30
ALMA-COMET	62.73	67.08	72.62	49.94	62.64	58.50	60.80	70.02	59.07	62.31	56.34
Ours-COMET	65.98	71.16	75.12	58.99	67.19	60.90	67.90	71.57	62.03	65.16	60.08
						N=200					
						Easy instances					
ALMA-BLEU	31.96	47.11	39.94	26.22	42.13	32.50	32.75	28.54	41.08	30.22	32.22
Ours-BLEU	33.10	47.41	41.60	25.79	41.43	31.52	32.20	26.91	40.48	29.79	32.04
ALMA-COMET	86.34	88.61	88.72	82.31	87.02	87.76	91.85	88.67	87.97	89.67	78.92
Ours-COMET	86.16	88.40	88.43	81.98	86.89	86.75	90.77	86.94	87.12	88.73	78.34
						Hard instances					
ALMA-BLEU	17.46	30.39	24.17	6.00	20.03	19.11	14.83	19.02	18.61	15.43	16.96
Ours-BLEU	19.31	35.04	29.25	7.55	23.70	19.96	18.16	18.29	21.52	15.95	19.28
ALMA-COMET	67.24	71.82	76.62	57.84	67.59	64.30	67.13	74.56	65.46	67.59	61.26
Ours-COMET	69.85	74.82	78.52	63.87	70.22	66.77	70.37	74.13	67.50	68.78	63.60
						N=500					
						Easy instances					
ALMA-BLEU	34.36	50.81	46.92	28.50	45.16	34.61	35.28	31.79	43.91	32.13	35.13
Ours-BLEU	35.33	50.59	47.25	27.82	44.16	33.25	34.07	30.00	42.92	31.67	34.54
ALMA-COMET	88.08	90.54	91.04	84.29	88.62	89.59	93.66	91.08	89.79	91.47	80.67
Ours-COMET	87.80	90.10	90.50	83.86	88.40	88.55	92.48	89.57	88.79	90.61	80.00
						Hard instances					
ALMA-BLEU	21.31	35.46	28.66	13.08	25.4	22.53	19.82	22.52	24.81	19.78	21.36
Ours-BLEU	23.09	37.91	32.66	14.04	27.32	22.89	22.38	21.32	26.58	19.78	22.82
ALMA-COMET	73.56	78.24	81.55	67.07	74.39	72.74	76.38	80.61	73.38	75.29	67.79
Ours-COMET	74.77	79.75	82.41	69.56	75.63	73.24	77.34	79.19	74.12	74.97	68.60
						Overall performance, i.e., N=infinite when all instances are included.					
ALMA-BLEU	30.73	44.68	36.46	24.15	40.37	31.37	31.12	26.67	39.05	28.76	30.46
Ours-BLEU	31.93	45.60	38.85	23.94	40.09	30.64	30.91	25.22	38.76	28.43	30.59
ALMA-COMET	84.42	86.29	86.30	79.70	85.09	85.45	89.42	85.85	85.76	87.50	76.83
Ours-COMET	84.50	86.53	86.45	80.05	85.22	84.78	88.75	84.38	85.19	86.77	76.59

Table 6: Specific results on 10 translation directions. The size of models are 13B. BLEU and COMET are reported. Cells where the difference is larger than 1.0 are highlighted with colored background. Blue indicates ours model outperforms ALMA and red indicates the opposite.