Mitigating Hallucinated Translations in Large Language Models with Hallucination-focused Preference Optimization

Anonymous ACL submission

Abstract

Machine Translation (MT) is undergoing a paradigm shift, with systems based on fine-tuned large language models (LLM) becoming increasingly competitive with traditional encoder-decoder models trained specifically for translation tasks. However, LLM-based systems are at a higher risk of generating hallucinations, which can severely undermine user's trust and safety. Most prior research on hallucination mitigation focuses on traditional MT models, with solutions that involve *post-hoc* mitigation − detecting hallucinated translations and re-translating them. While effective, this approach introduces additional complexity in deploying extra tools in production and also increases latency. To address these limitations, we propose a method that intrinsically learns to mitigate hallucinations during the model training phase. Specifically, we introduce a data creation framework to generate hallucination focused preference datasets. Fine-tuning LLMs on these preference datasets reduces the hallucination rate by an average of 96% across five language pairs, while preserving overall translation quality. In a zero-shot setting our approach reduces hallucinations by 89% on an average across three unseen target languages.

1 Introduction

LLMs are gaining popularity for various NLP applications, including machine translation. Fine-tuning LLMs for MT has been proven to be highly data-efficient, requiring orders of magnitude less parallel data than large standalone multilingual MT models, while achieving increasingly competitive performance (Liao et al., 2024; Xu et al., 2024; Alves et al., 2024). Moreover, there is a significant and ongoing effort within the research community to push the performance limits of foundational LLMs and expand their multilingual capabilities (Jiang et al., 2023; Dubey et al., 2024; Aryabumi et al., 2024).

Despite these advantages, LLM-based models are more prone to *hallucinations*: the models generates information that is inaccurate or entirely fabricated. This issue has lead to a growing research area, focusing on the causes, detection, and mitigation of hallucinations in LLMs (Tonmoy et al., 2024). In the context of MT, hallucinations manifest as highly pathological translations, which can lead to misunderstandings in conversations, potentially damaging relationships and undermining user trust in the system (Kumar et al., 2023).

Most of the existing research on hallucination mitigation in MT has focused on traditional encode-decoder models, establishing effective post-hoc mitigation strategies (Guerreiro et al., 2023c; Dale et al., 2023a,b). These strategies first detect whether a translation contains hallucination, and if so, generate and present a *mitigated* translation to the user. In practical scenarios, using post-hoc mitigation has several drawbacks: i) the need for deploying an additional hallucination detector in production; ii) running the hallucination detector on every translation, which increases cost and latency; and iii) re-running inference if a translation hallucinates (which often much slower than regular inference).

To address these issues, we propose a framework that intrinsically integrates hallucination mitigation during the LLM development phase, aiming to minimize hallucinations from the outset. Specifically, we apply post-hoc mitigation strategies *offline* on a large-scale monolingual corpus, generating a corpus of model hallucinations alongside their corresponding mitigated translations. We then fine-tune the LLM using Contrastive Preference Optimization (CPO) (Xu et al., 2024), guiding the model away from hallucinations.

Our approach requires no additional human-annotated data, is easily scalable across many language pairs, and is highly effective − achieving a 96% reduction in hallucination rates across five

language pairs without sacrificing general translation quality. It also generalizes well, achieving an average 89% reduction in hallucination rates across three unseen target languages. Overall, our main contributions include:

- Proposing a novel approach for creating hallucination-focused preference datasets.
- Identifying the most effective fine-tuning technique for leveraging this preference dataset.
- Exploring the cross-lingual generalization capabilities of the fine-tuned models in a zero-shot setting.
- Determining the most effective post-hoc mitigation strategies for LLM based translation models.

2 Dataset Creation Framework

One of the techniques for fine-tuning LLMs for translation is preference optimization (Xu et al., 2024) which uses a dataset of triplets, consisting of a source sentence x, its preferred translation y_p, and a dispreferred translation y_d. Preference optimization trains the model to prioritize the generation of preferred set of translations over dispreferred ones. Xu et al. (2024) focus on optimizing general translation quality, and hence in their datasets, y_p and y_d differ only in quality and do not explicitly consider the notion of hallucination. For instance, both translations could be broadly correct, but one might be preferred over the other due to minor errors or subtle differences in style.

To address hallucinations, we develop a framework for automatically creating a *hallucination focused* preference dataset and propose to fine-tune the LLM on this dataset to effectively mitigate hallucination generation. In this dataset, the dispreferred translations contain hallucinations, whereas the preferred translations do not. The set of dispreferred translations are derived from the LLM's own generated outputs. This is particularly important as it enables the model to learn from its own errors and correct them. Our approach for creating this preference dataset is completely unsupervised and can easily scale to multiple languages without any human annotation. At a high level, the dataset creation process consists of translating monolingual data using the LLM and automatically detecting hallucinations (Section 2.1) and mitigating them using existing post-hoc methods (Section 2.2)

2.1 Hallucination Detection

In the first step, we construct a set of source sentences and their corresponding dispreferred translations containing hallucinations. To achieve this, we translate publicly available monolingual corpora $\mathcal{D}_m$ from the source language into the target languages using the model $\mathcal{M}$, which we aim to fine-tune for reducing hallucinations. We then automatically identify translations y ($y := \mathcal{M}(x)$) that exhibit hallucination using the state-of-the-art hallucination detector model based on `BLASER 2.0-QE` (Chen et al., 2023; Dale et al., 2023b). `BLASER 2.0-QE` is a reference-free machine translation quality estimation metric that predicts cross-lingual semantic similarity between a source sentence x its translation y. It operates on a scale of 1-5, where 1 denotes completely unrelated sentences and 5 signifies fully semantically equivalent sentences. We re-normalize the `BLASER` score to a hallucination score (`HS`), with a higher value indicating a greater likelihood of hallucination in y:

$$\mathsf{HS}(x, y) = 1 - \frac{\mathsf{BLASER}(x, y)}{5} \tag{1}$$

After fixing a threshold T, we classify a translation as containing hallucination if its hallucination score exceeds the threshold. Collecting such instances where hallucinations are detected provides us with a hallucination dataset $\mathcal{D}_h$, which consists of source sentences and their corresponding hallucinated translations as follows:

$$\mathcal{D}_h := \{(x, y) \mid \mathsf{HS}(x, y) \geq T \; \forall \, x \in \mathcal{D}_m\} \tag{2}$$

2.2 Post-hoc Hallucination Mitigation

The second step involves mitigating the hallucinated translations in $\mathcal{D}_h$ to create hallucination-free alternatives. Previous works (Dale et al., 2023a; Guerreiro et al., 2023a,c) have proposed several post-hoc mitigation strategies, though they are typically applied during test time. In contrast, we explore using these strategies offline to build a preference fine-tuning corpus. We consider a few notable strategies, outlined below:

Fallback System Guerreiro et al. (2023a) demonstrated that simply switching to a different fallback translation system when hallucinations occur is an effective mitigation strategy. Following this, we employ the `NLLB-3.3B` model (NLLB Team et al., 2022) as a fallback.

Candidate Generation and Selection Dale et al. (2023a) propose generating multiple alternative translation candidates from the original model and selecting one of them as the mitigated translation based on a specific criterion. This approach involves two degrees of freedom: (i) candidate generation and (ii) candidate selection. To generate n candidates we explore the following strategies:

- **MC beam**: Using n iterations of beam search with Monte Carlo dropout (Gal and Ghahramani, 2016).
- **Temperature sampling**: Sampling from the full probability distribution, adjusted by a temperature parameter t, to control the sharpness of the distribution.
- **Nucleus sampling**: Sampling from a set of tokens that covers top $p\%$ of the posterior probability distribution at each step (Holtzman et al., 2020).
- **Epsilon sampling**: Sampling from a set of tokens where each token has a probability greater than or equal to a threshold ϵ (Hewitt et al., 2022; Freitag et al., 2023).

To select the best candidate, we explore the following algorithms:

- **MBR decoding**: Selects the candidate that maximizes the average utility with respect to all other candidates. (Kumar and Byrne, 2004; Freitag et al., 2022). We evaluate utility between two candidates using `chrF` (Popovic, 2015), `LaBSE` (Feng et al., 2022), and `COMET` (Rei et al., 2022).
- **Re-ranking**: Selects the candidate that maximizes utility with respect to the source sentence, using `LaBSE` and `COMET` (Rei et al., 2020) as utility metrics.

We compare the effectiveness these strategies in mitigating hallucinations, analyzing the impact of different generation and sampling methods in Section 5.

We select the best mitigation strategy based on a held out development set and use it to generate alternative translations $\tilde{y}$ corresponding to each sample $(x, y) \in \mathcal{D}_h$. We construct our hallucination focused preference fine-tuning dataset $\mathcal{D}_p$ by retaining samples where the alternative translation successfully mitigates hallucination ($\mathtt{HS}(x, \tilde{y}) < T$). Formally $\mathcal{D}_p$ is defined as follows:

$$\mathcal{D}_p := \{(x, \tilde{y}, y) \mid \mathtt{HS}(x, \tilde{y}) < T \; \forall \, (x, y) \in \mathcal{D}_h\} \quad (3)$$

3 Fine-tuning Using CPO

We fine-tune the baseline LLM $\mathcal{M}$ using our hallucination-focused preference dataset $\mathcal{D}_p$ through CPO, a variant of Direct Preference Optimization (DPO) (Rafailov et al., 2023), which has shown to be effective for fine-tuning LLMs on the translation task. The CPO objective is formally defined as follows:

$$\mathcal{L}_{CPO} = \mathcal{L}_{NLL} + \mathcal{L}_P \quad (4)$$

where

$$\mathcal{L}_P = -\,\mathbb{E}_{(x,y_p,y_d)\sim\mathcal{D}_p} \log \sigma\big(\beta \log \frac{\pi_\theta(y_p|x)}{\pi_\theta(y_d|x)}\big) \quad (5)$$

$$\mathcal{L}_{NLL} = -\,\mathbb{E}_{(x,y_p,y_d)\sim\mathcal{D}_p} \log \pi_\theta(y_p|x) \quad (6)$$

In equations above, x, y_p and y_d represent the source sentence, preferred (hallucination free) translation and dispreferred (hallucination containing) translation, respectively, sampled from the preference dataset $\mathcal{D}_p$. The policy π_θ refers to the conditional probability distribution from the model $\mathcal{M}$, σ is the sigmoid function and β is a scaling hyperparameter from (Rafailov et al., 2023).

The CPO objective combines the standard negative log-likelihood `NLL` loss, which encourages the model to generate y_p, and the preference loss $\mathcal{L}_p$, which aims to increase the probability gap between y_p and y_d. The preference loss term explicitly instructs the model to prioritize the generation of y_p and reject y_d. In Section 6.2 we show that this loss term is crucial for reducing the model's likelihood of generating hallucinations.

In our dataset, we ensure that y_p always has higher quality than y_d, as measured by hallucination score ($\mathtt{HS}(x, y_p) < T$ and $\mathtt{HS}(x, y_d) \geq T$). However different preference pairs may exhibit varying quality gaps. To account for this variation in quality gaps in the preference fine-tuning, we introduce a scaling term to $\mathcal{L}_P$. A preference pair (y_p, y_d) with larger quality gap provides a more informative data point, so we design the scaling term to assign greater weight to pairs with a larger gaps, proportional to the quality ratio of y_p and y_d. With this scaling term, the modified preference loss ($\mathcal{L}'_p$) is defined as follows[1]:

[1]We found scaled CPO performs slightly better then standard CPO as shown in Table 10 in Appendix C.

$$\mathcal{L}'_p = -\mathbb{E}_{(x,y_p,y_d)\sim\mathcal{D}_p}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_p|x)}{\pi_\theta(y_d|x)} + \beta\log\frac{\phi(x,y_p)}{\phi(x,y_d)}\right)\right] \quad (7)$$

where, ϕ is a scoring function that measures the quality of a translation given the source. We choose ϕ to be the hallucination score (HS). With this change, our final CPO loss is shown in equation 8

$$\mathcal{L}'_{CPO} = \mathcal{L}'_p + \mathcal{L}_{NLL} \quad (8)$$

4 Experimental Setup

4.1 Evaluation Metrics

Given a model $\mathcal{M}$, we evaluate it on a monolingual dataset $\mathcal{D}$ using *hallucination rate*. Hallucination rate (HR) computes the ratio of source sentences for which model produces translations containing hallucinations:

$$\mathrm{HR}(\mathcal{M},\mathcal{D}) = \frac{|\{x \mid \mathrm{HS}(x,\mathcal{M}(x)) \geq T \;\; \forall x \in \mathcal{D}\}|}{|\mathcal{D}|} \quad (9)$$

where $|\cdot|$ counts the number of elements in a set.

We split the monolingual corpus $\mathcal{D}_m$ into $\mathcal{D}_m^{train}$ (train), $\mathcal{D}_m^{dev}$ (dev) and $\mathcal{D}_m^{test}$ (test) sets. The hallucination-focused preference dataset ($\mathcal{D}_p$) is derived from $\mathcal{D}_m^{train}$ as described as Section 2. We evaluate the baseline and fine-tuned LLMs using hallucination rates computed against unseen set $\mathcal{D}_m^{test}$. All the hyperparameters and the best post-hoc mitigation strategy for preparing the fine-tuning set are selected based on $\mathcal{D}_m^{dev}$.

To ensure that improvements in hallucination mitigation do not come at the expense of general translation quality, we also evaluate the baseline and fine-tuned models on the WMT'22 and WMT'23 testsets using three COMET models: `wmt22-cometkiwi-da`, `wmt23-cometkiwi-da-xxl`, and `XCOMET-XXL`. This evaluation methodology aligns with that of Xu et al. (2024).

4.2 Baseline Model and Language Coverage

We choose `ALMA-7B-R` as our baseline LLM. Built upon `LLAMA-2` (Touvron et al., 2023), `ALMA-7B-R` has been extensively optimized for translation through multiple rounds of fine-tuning, including continued pre-training on multilingual data, supervised fine-tuning with parallel corpora and preference tuning using CPO. `ALMA-7B-R` has shown competitive performance, matching or surpassing top systems in WMT shared evaluation, and even GPT-4 (OpenAI et al., 2024), making it a strong baseline for our hallucination mitigation experiments.

`ALMA-7B-R` supports translation across ten language directions: English↔{Czech (*cs*), German (*de*), Icelandic (*is*), Russian (*ru*) and Chinese (*zh*)}. However due to resource constraints, in our study, we focus on a subset of five language pairs: *en*→{*cs*, *de*, *is*, *ru*, *zh*}.

4.3 Hallucination Focused Preference Dataset Construction

We follow the data creation framework outlined in Section 2 to construct a hallucination focused preference fine-tuning dataset, as detailed below:

4.3.1 Monolingual Data

As our study is restricted to language pairs with English as source, we randomly sample English sentences from the NewsCrawl dataset (Kocmi et al., 2022)[2] for $\mathcal{D}_m$. We sample 0.5M sentences each for $\mathcal{D}_m^{dev}$ and $\mathcal{D}_m^{test}$, and these evaluation sets are shared across all language pairs. To create preference sets for each language pair, we sample separate $\mathcal{D}_m^{train}$ sets, with sizes of 2M (*en*→*zh*), 5M (*en*→*cs*, *en*→*is*, *en*→*ru*), or 10M (*en*→*de*) sentences. The sizes are determined based on hallucination rates of the baseline model for each language pair, with larger sets allocated to language pairs exhibiting lower hallucination rates, ensuring that the resulting preference sets are of comparable sizes across all language pairs. All the above datasets are cleaned by applying a series of filters to remove noisy samples.[3]

4.3.2 Hallucination Detection

As outlined in Section 2.1, for each language pair, we translate the corresponding $\mathcal{D}_m^{train}, \mathcal{D}_m^{dev}, \mathcal{D}_m^{test}$ sets using the baseline `ALMA-7B-R` into the target language. We then create the corresponding hallucination datasets $\mathcal{D}_h^{train}, \mathcal{D}_h^{dev}, \mathcal{D}_h^{test}$ by retaining translations where hallucination score exceeds the threshold T. We set T to be 0.5 based on manual verification of the resulting $\mathcal{D}_h^{dev}$ sets for *en*→*zh* and *en*→*de*. Native chinese and german speakers verified that 97% and 87% of translations in the *en*→*zh* and *en*→*de* sets, respectively, did contain highly pathological errors. Consequently, this

[2] https://data.statmt.org/news-crawl/ (2023 release)

[3] Appendix A provides more information on the filtering process, and monolingual data statistics.

threshold is adopted for all language pairs throughout our study, unless otherwise specified. The number of samples in hallucination datasets for each split and language pair, along with the corresponding hallucination rates (%) are summarized in Table 1.

	$\mathcal{D}_h^{train}$	$\mathcal{D}_h^{dev}$	$\mathcal{D}_h^{test}$
en→cs	2085 (0.04)	202 (0.04)	179 (0.04)
en→de	673 (0.01)	47 (0.01)	39 (0.01)
en→is	3682 (0.08)	384 (0.08)	388 (0.08)
en→ru	1933 (0.04)	186 (0.04)	196 (0.04)
en→zh	8470 (0.45)	2178 (0.46)	2192 (0.46)

Table 1: Hallucination count (HR in %) for ALMA-7B-R.

4.3.3 Post-hoc Hallucination Mitigation

We evaluate the post-hoc mitigation strategies described in Section 2.2 on $\mathcal{D}_h^{dev}$. Given a sample $(x, y_d) \in \mathcal{D}_h^{dev}$, where y_d contains hallucinations, each mitigation strategy $\mathcal{S}$ attempts to generate an alternative translation $\tilde{y} := \mathcal{S}(x)$ which is likely free of hallucinations. We evaluate these strategies using *mitigation rate* (MR), which is the ratio of samples where $\tilde{y}$ successfully mitigates hallucinations. Higher MR values indicate better performance.

$$\text{MR}(\mathcal{S}, \mathcal{D}_h) = \frac{|\{x \mid \text{HS}(x, \mathcal{S}(x)) < T \;\; \forall (x, y_d) \in \mathcal{D}_h\}|}{|\mathcal{D}_h|} \tag{10}$$

For the *Fallback* strategy, we use a beam size of 40. For *Candidate Generation and Selection* approach, we generate $n = 40$ candidates using temperature sampling with $t \in \{0.8, 1, 1.5\}$ in conjunction with either nucleus sampling with $p = 0.9$ or epsilon sampling with $\epsilon = 0.02$. For MCBeam, we generate candidates using a beam size of 5. When using COMET with MBR, we use eamt22-cometinho-da which is a distilled model that takes as input the source sentence, translation and reference translation. For COMET with Re-ranking, we employ the wmt20-comet-qe-da, which only takes the source sentence and translation as input.

A detailed comparison of the mitigation strategies is presented in Section 5.1. We use the best performing strategy to construct our preferences datasets $\mathcal{D}_p^{train}$ from $\mathcal{D}_h^{train}$ as described in Section 2.2. The number of samples in these preference datasets across all language pairs is presented in Table 2.[4]

en→cs	*en→de*	*en→is*	*en→ru*	*en→zh*
2063	671	3598	1931	8349

Table 2: Number of samples in $\mathcal{D}_p^{train}$.

4.4 Combining Hallucination and Translation Quality Preference Datasets

While $\mathcal{D}_p^{train}$ is specifically constructed to mitigate hallucinations, fine-tuning solely on this dataset can lead to a decline in general translation quality. To address this, we mix $\mathcal{D}_p^{train}$ with $\mathcal{D}_{alma}^{train}$, the preference dataset originally used to fine-tune the baseline ALMA-R model by Xu et al. (2024), which focuses on overall translation quality. Combining the two sets helps preserve the original translation quality while improving hallucination mitigation. $\mathcal{D}_{alma}^{train}$ is comparable in size to $\mathcal{D}_p^{train}$, with detailed statistics provided in Table 18

4.5 Fine-tuning Using CPO

We adhere to a fine-tuning setup that closely follows the methodology described in Xu et al. (2024). In line with their approach, we fine-tune LoRA adapters (Hu et al., 2022) and utilize the same prompt structure.[5] For the modified preference loss function ($\mathcal{L'}_p$), we use HS as the scoring function ϕ for $\mathcal{D}_p^{train}$ and COMET as the scoring function for the $\mathcal{D}_{alma}^{train}$ dataset.[6] We normalize the scoring functions of both datasets to ensure their ranges align with each other.

Most hyper-parameters are optimized based on the hallucination rate of the fine-tuned model on the smaller $\mathcal{D}_h^{dev}$ sets to facilitate quick iterations. However, when multiple configurations yield similar results, we decide based on the full development set D_m^{dev}.

5 Results

We present the comparison between different post-hoc mitigation strategies in Section 5.1 and main results of our fine-tuned models in Section 5.2

5.1 Post-hoc Mitigation Strategies

Table 3 summarizes the mitigation rates of various strategies across different selection methods and utility metrics, focusing on the top performing sampling settings.[7] All strategies significantly

[4] Detailed statistics comparing the hallucination scores and lengths of preferred vs. dispreferred translations can be found in Appendix, Section G

[5] More details on hyperparameters are available in Appendix B.

[6] The COMET scores are part of the original preference dataset, which are average of KIKI-XXL and XCOMET.

[7] Table 11 in Appendix shows several sampling methods.

	Fallback	MBR					Re-ranking			
	NLLB-3.3B	chrf	COMET	LaBSE	COMET $\epsilon = 0.02$	LaBSE $\epsilon = 0.02$	COMET	LaBSE	COMET $\epsilon = 0.02$	LaBSE $\epsilon = 0.02$
en→cs	**100**	96.6	96.1	97.6	97.6	97.1	98.1	99.5	98.1	99.5
en→de	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
en→is	98.3	92.3	92.9	95.4	95.1	95.4	95.7	97.7	96.3	**98.9**
en→ru	97.4	99.0	99.5	98.4	98.4	99.5	99.0	99.5	**100**	**100**
en→zh	86.9	97.6	98.1	98.4	98.6	99.1	96.9	99.1	97.1	**99.4**
Average	96.5	97.1	97.3	98.0	97.9	98.2	97.9	99.2	98.3	**99.6**

Table 3: Mitigation rates MR in % (↑) for different post-hoc mitigation strategies on $\mathcal{D}_h^{dev}$ set.

Model	Hallucination count/rate (↓)							WMT'23 COMET(↑)
	en→cs	*en→de*	*en→is*	*en→ru*	*en→zh*	avg	avg HR (%)	*en→X*
NLLB-3.3B	471	732	1459	252	38302	8243	1.743	75.9
ALMA-7B-R	179	39	388	196	2192	599	0.127	**81.8**
$\mathcal{M}_p$	5	2	37	2	**74**	**24**	**0.005**	80.8
$\mathcal{M}_{p+a}$	**4**	**1**	**35**	**0**	80	**24**	**0.005**	81.6
ALMA-7B-R + *post-hoc**	1	0	8	1	28	7.6	0.002	-

Table 4: Main Results: Hallucination count and HR (%) on $\mathcal{D}_m^{test}$, and average COMET scores on WMT'23 testsets. * indicates an upper bound and should be seen as a reference point since it is not a modeling technique.

reduce hallucinations, with even the worst performing one achieving an average mitigation rate of over 96%. The optimal setting, achieved through epsilon sampling with $\epsilon = 0.02$ followed by re-ranking with LaBSE, results in an impressive average mitigation rate of 99.6%. Notably, we observe that for both MBR and Re-rank, LaBSE consistently outperforms COMET. This aligns with previous research on hallucination detection, which has shown LaBSE to be superior to COMET (Dale et al., 2023b). Furthermore, model-based metrics for MBR, such as COMET and LaBSE outperform chrF. Comparing both candidate selection methods overall, Re-rank outperforms MBR. The Fallback strategy using NLLB-3.3B achieves a mitigation rate of 96.5%. While quite substantial, it falls short of the best results, possibly due to the baseline ALMA-7B-R being a stronger model, generating higher quality and more diverse translations.[8]

5.2 Fine-tuning Using CPO

We present the main results in Table 4.[9] Our primary baseline, ALMA-7B-R, achieves an average hallucination rate of 0.127%. ALMA-7B-R is a much stronger baseline compared to traditional encoder-decoder based NLLB-3.3B, which exhibits an average hallucination rate of 1.73%, nearly 14 times higher than that of ALMA-7B-R. This difference is expected, given that ALMA-7B-R is a stronger translation model, as reflected by its superior average COMET scores on the WMT testsets. Examining the hallucination rates across all language pairs, we observe that the *en→zh* language pair consistently shows the highest hallucination rates across all the models.

Next, we analyze the results obtained by fine-tuning ALMA-7B-R on different preference datasets. Fine-tuning using our hallucination focused preference dataset $\mathcal{D}_p^{train}$, gives us model $\mathcal{M}_p$. The hallucination rate of this model drops significantly from 0.127% to an average of 0.005%. This demonstrates the effectiveness of our unsupervised preference data creation approach, resulting in a remarkable 96% reduction. However, along the reduction in hallucinations, we observe a decline in general translation quality, with the average COMET score dropping by 1.0 from the baseline.

To mitigate this drop in translation quality, we fine-tune the model using a combined dataset $\mathcal{D}_p^{train} \cup \mathcal{D}_{alma}^{train}$, which gives us model $\mathcal{M}_{p+a}$. By balancing training between hallucination mitigation and general translation tasks, we observe an improvement of 0.8 points in the average COMET score, bringing the model nearly on par with the baseline performance, while still maintaining hallucination rate of 0.005%. For more detailed general translation quality comparisons, refer to Section L in the Appendix. To establish an upper bound, we apply the best post-hoc mitigation strategy to the hallucinations from the baseline ALMA-7B-R model, reporting this as ALMA-7B-R + *post-hoc* in Table 4. This represents using the post-hoc mitigation sys-

[8] Comparison of NNLB-3.3B and ALMA-7B-R on general translation quality is shown in Table 21, 22 in Appendix.

[9] Individual COMET model scores for WMT'22 and WMT'23 across each language pair are detailed in Table 21, 22, 23, 24

tem during test time. Our findings indicate that our best model, with a hallucination rate of 0.005%, comes very close to this upper bound of 0.002%, without requiring any additional mitigation systems at test time.

6 Analysis and Discussions

6.1 Cross-lingual Zero-shot Generalization

To assess the cross-lingual generalization of our fine-tuning approach in reducing hallucinations on unseen language pairs, we conducted zero-shot experiments comparing baseline ALMA-7B-R with our best fine-tuned model ($\mathcal{M}_{p+a}$) in a zero-shot setting. In these experiments, we translated our test set $\mathcal{D}_m^{test}$ from English into three target languages – French (*fr*), Italian (*it*), and Spanish (*es*), none of which were prominently present in the pre-training and fine-tuning stages of ALMA-7B-R.

Table 5 presents the hallucination rates and COMET scores for both models across these language pairs.[10] Notably, both models perform well, despite the target languages being unseen during training. The baseline model achieves an average COMET score of 83.31, with the fine-tuned model trailing slightly at 83.17. However, in terms of hallucination rates, the fine-tuned model significantly outperforms the baseline, reducing the average hallucination rate from 0.273% to 0.03%, representing an 89% reduction. These results demonstrate that our fine-tuning approach generalizes effectively to unseen language pairs, substantially reducing hallucinations without significant loss in general translation quality.

	HR % (↓)		COMET (↑)	
	ALMA-7B-R	$\mathcal{M}_{p+a}$	ALMA-7B-R	$\mathcal{M}_{p+a}$
en→es	0.164	0.007	83.30	83.25
en→fr	0.399	0.077	83.05	82.39
en→it	0.256	0.007	83.57	83.87
Average	0.273	0.030	83.31	83.17

Table 5: Cross-lingual zero-shot results.

6.2 Ablation of Loss Function Components

As shown in equation 8, the CPO loss consists of two components: i) preference loss and ii) NLL loss. We conducted an ablation study to understand the contribution of each component. When only the NLL loss is active, it corresponds to supervised fine-tuning (SFT) on the source and mitigated translations. We optimized the hyperparameters corresponding for each loss configuration based on hallucination rates on the $\mathcal{D}_h^{dev}$ set and then evaluated both the baseline and the fine-tuned models on the full $\mathcal{D}_m^{dev}$ set.

Table 6 summarizes the results of these ablations. The findings reveal that using only the preference loss results in poor performance, with a hallucination rate of 3.556%, which is significantly worse than the baseline ALMA-7B-R (0.127%). In contrast, using only the NLL loss yields a lower hallucination rate of 0.078%, outperforming the baseline. However, the best performance is achieved when both losses are combined, reducing the hallucination rate to just 0.005%. This demonstrates the effectiveness of the CPO loss over simple SFT using mitigated translations, highlighting the complementary benefits of preference and cross-entropy losses.

	ALMA-7B-R	$\mathcal{M}_p$		
		$\mathcal{L}'_P$	$\mathcal{L}_{NLL}$	$\mathcal{L}'_{CPO}$
en→cs	202	10	216	7
en→de	47	5	124	1
en→is	384	72	441	37
en→ru	186	83836	127	2
en→zh	2178	174	931	73
Avg. HR (%)	0.127	3.556	0.078	0.005

Table 6: Hallucination counts (HR in %) of ALMA-7B-R and $\mathcal{M}_p$ using different loss variants on $\mathcal{D}_m^{dev}$.

6.3 Ablation of Data Quantity vs. Quality

To create $\mathcal{D}_p^{train}$, we select dispreferred translations with a hallucination score ≥ 0.5. Lowering this threshold yield more training samples, but risks including translations that do not accurately reflect true hallucinations, thus reducing the quality of the preference dataset. To explore the tradeoff between data quantity and quality, we conducted an experiment by creating a version of $\mathcal{D}_p^{train}$ with a lower threshold of 0.45. We fine-tuneed the baseline on both versions of the preference dataset and evaluated the models on $\mathcal{D}_m^{dev}$. As shown in Table 7, lowering the threshold to increase the dataset size led to a decline in performance, indicating that the quality of the preference data is more crucial than its quantity.

6.4 Distribution of Hallucination Scores

Figure 1 illustrates the distribution of hallucination scores for the *en→zh* pair on $\mathcal{D}_m^{test}$ before and after fine-tuning. The top plot shows the full scale distribution from 0-1, while the bottom image provides a

[10]The COMET scores were computed using the *wmt22-cometkiwi-da* model.

	0.5 (default)	0.45
en→cs	7	24
en→de	1	5
en→is	37	135
en→ru	2	10
en→zh	73	259
Avg. rate (%)	0.005	0.018

Table 7: Hallucination counts (HR in %) on $\mathcal{D}_m^{dev}$ after fine-tuning with $\mathcal{D}_p^{train}$ collected at different thresholds.

zoomed-in view focused on the critical range of 0.5-1, which highlights the hallucination-prone section. In the top plot, the distribution post-fine-tuning (in orange) shifts markedly to the left, indicating an overall improvement in translation quality across the dataset. In the bottom plot, we observe that the remaining hallucinations post-fine-tuning are primarily concentrated near the threshold, with fewer instances with extreme hallucination scores. Plots for all language pairs can be found in Appendix 3.

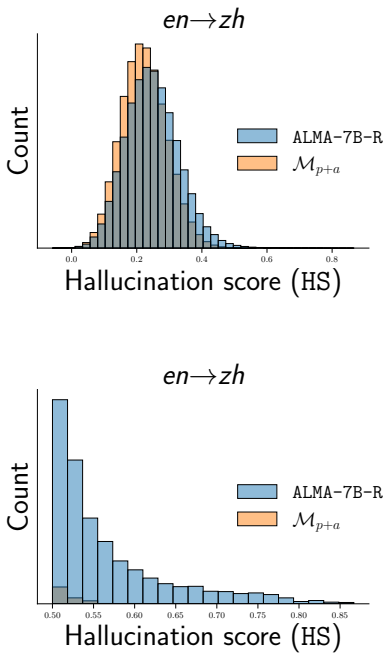

Figure 1: Distribution of the hallucination scores on $\mathcal{D}_m^{test}$ for ALMA-7B-R and $\mathcal{M}_{p+a}$.

7 Related Work

Prior works on hallucination detection include identifying repeated n-gram patterns in translations (Raunak et al., 2021), utilizing internal model information such as attention weights (Lee et al., 2019; Berard et al., 2019; Ferrando et al., 2022b,a; Voita et al., 2021; Xu et al., 2023; Guerreiro et al., 2023b), and estimating uncertainty using the model's sequence log-probability (Guerreiro et al., 2023c). Other works have explored external models based on quality estimation (COMET-QE) and cross-lingual sentence similarity (LASER, LaBSE, XNLI, BLASER-QE) (Dale et al., 2023a,b).

To mitigate hallucinations, prior works have primarily focused on *post-hoc* solutions. These include using a fallback model (Guerreiro et al., 2023a), generating multiple candidates and selecting the best using a re-ranker (Guerreiro et al., 2023c), or applying consensus-based decoding strategies such as Minimum Bayes Risk (MBR) (Eikema and Aziz, 2020). Other approaches have explored contrastive decoding by leveraging probabilities from different models (Li et al., 2023), using previous output tokens (Su and Collier, 2023), or utilizing a contrastive input (Sennrich et al., 2024). While all these approaches mitigate hallucinations during or after inference, our approach takes an orthogonal path by addressing the issue directly within the model itself.

8 Conclusion

In this work, we presented a framework for mitigating translation hallucinations in large language models (LLMs). To the best of our knowledge, this is among the first works to demonstrate how to mitigate translation hallucination in LLMs. In this framework, we propose an unsupervised method to create a hallucination-focused preference dataset, which is easily scalable across multiple languages. Fine-tuning LLMs using this dataset through preference optimization reduces hallucination rates by an average of 96%, while preserving general translation quality. Additionally, our method generalizes well in a cross-lingual zero-shot setting, achieving an 89% reduction in hallucination rates across three previously unseen target languages.

Limitations

- In this work we explored only *en→X* language pairs due to time and resource constraints. We leave the exploration of other directions as a future work.

- Since natural translation hallucination is very rare, we need to translate huge amount of monolingual data to create a reasonable amount of hallucination focused preference

dataset, thus making our approach time and compute intensive.

- Our approach depends on a hallucination detector. So, language pairs of interest must be supported by the detector, as well as some analysis might be required to decide hallucination detector threshold.

Ethics Statement

This work, in our knowledge, doesn't pose any ethical concerns. It proposes approaches to make AI models safe and trustworthy. Still, our models might generate some hallucinations like any other AI models. The original data, model, tools, and open-source software used in the paper are publicly available and has been mentioned in the corresponding sections.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4506–4520. International Committee on Computational Linguistics.

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.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022a. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023b. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023c. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. 2024. Ikun for wmt24 general mt task: Llms are here for multilingual machine translation.

Marta R. Costa-jussà NLLB Team, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation. *Trans. Mach. Learn. Res.*, 2023.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55204–55224. PMLR.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

A Monolingual Data Filtering

To prepare the monolingual data for translation, we apply the following four filters in sequence. Table 8 shows the statistics of monolingual data before and after applying the filters.

Heuristic filter removes empty lines, replaces '\n' with '<NEWLINE>', eliminates sentences containing unprintable unicode characters, as well as those with Chinese decoding errors, and excludes rows with HTML or JSON-like elements.

Length filter splits the sentence by whitespace (since the source language is English), and removes sentences that are shorter than 5 words or longer than 100 words.

Deduplication filter removes exact duplication with `drop_duplicates` function from `Pandas` library[11].

Language ID filter identifies the language of each sentence using the `fasttext` model (Joulin et al., 2017, 2016) and removes sentences that fall below the language probability threshold of 0.5.

		Before filtering	After filtering
$\mathcal{D}_m^{dev}$	*en→X*	500K	473K
$\mathcal{D}_m^{test}$	*en→X*	500K	473K
$\mathcal{D}_m^{train}$	*en→cs*	5M	4.73M
	en→de	10M	9.46M
	en→is	5M	4.73M
	en→ru	5M	4.73M
	en→zh	2M	1.89M

Table 8: Monolingual data statistics.

B Hyperparameters for Fine-tuning Using CPO

For the preference fine-tuning process, we only train the LoRA parameters, specifically targeting *down_proj*, *q_proj*, *k_proj*, and *v_proj* with a rank of 16. We set the maximum sequence length to 768 tokens, utilize the Hugging Face accelerator with Fully Sharded Data Parallel (FSDP), and train on eight H100 GPUs, typically completing training in less than an hour. Inferences are performed on V100s, and takes roughly 7 GPU hours on $\mathcal{D}_h^{dev/test}$ and 1150 GPU hours on $\mathcal{D}_m^{dev/test}$. The value of β is set to 0.1, consistent with the findings of Rafailov et al. (2023) and Xu et al. (2024). We conduct a partial grid search for hyperparameters, varying the *batch size* from $\{16, 32, 64, 128, 256, 512\}$ and the *learning rate* from $\{2e{-}5, 5e{-}5, 1e{-}4, 2e{-}4, 5e{-}4\}$. Through our experimentation, we find that setting epoch to 1 generally suffices for optimal performance. We use beam size of 5 for baseline and all fine-tuned models.

The best hyperparameters we found for $\mathcal{M}_p$ and $\mathcal{M}_{p+a}$ are listed in Table 9

	$\mathcal{M}_p$	$\mathcal{M}_{p+a}$
batch size	16	128
learning rate	$1e-4$	$5e-4$
scheduler	inverse_sqrt	inverse_sqrt
optimizer	AdamW	AdamW
epoch	1	1
β	0.1	0.1

Table 9: Best hyperparameters found on $\mathcal{D}_h^{dev}$ for the model $\mathcal{M}_p$ and $\mathcal{M}_{p+a}$

C Standard CPO vs. Scaled CPO

We conducted an evaluation on $\mathcal{D}_h^{dev}$ to compare the performance of standard ($\mathcal{L}_{CPO}$) vs. scaled CPO ($\mathcal{L}'_{CPO}$) losses. Our results show that $\mathcal{L}'_{CPO}$ achieves an average hallucination rate of 0.774%, outperforming $\mathcal{L}_{CPO}$, which has an average rate of 1.028%. Table 10 presents a comparison of the two methods across all five language pairs.

	$\mathcal{L}_{CPO}$	$\mathcal{L}'_{CPO}$
en→cs	2.475	0.990
en→de	0.000	0.000
en→is	1.837	2.100
en→ru	0.000	0.000
en→zh	0.827	0.781
Average	1.028	0.774

Table 10: Hallucination rate HR (%) on $\mathcal{D}_h^{dev}$ for the model $\mathcal{M}_p$ fine-tuned with different CPO loss variants.

D Hallucination Characterization

To gain a deeper understanding of the nature of hallucinations, we conducted a detailed analysis of the source sentences and the corresponding hallucinated translations on the test set $\mathcal{D}_h^{test}$

Source sentences We examined source sentences to identify any patterns that might consistently trigger hallucinations when translating to different target languages. Table 13 presents these statistics of the overlap of source sentences between hallucination samples of different language pairs. For e.g., in the *en→zh* language pair (first row), 2178 source sentences generate hallucinations, however only

[11] https://pandas.pydata.org

	Fallback	MBR							Re-rank						
	NLLB	chrF	COMET				LaBSE		COMET			LaBSE			
	Beam	$t=1$	$t=1$	$t=1.5$	$t=1$	$t=2.0$	$t=1$	$t=1$	$t=1$	$t=1$	$t=0.8$	$t=1$	$t=1$	$t=1$	$t=0.8$
				$p=0.9$	$\epsilon=0.02$	$\epsilon=0.02$		$\epsilon=0.02$		$\epsilon=0.02$			MCbeam	$\epsilon=0.02$	$\epsilon=0.02$
en→cs	**100**	96.6	96.1	96.6	97.6	97.1	97.6	97.1	98.1	98.1	96.6	99.5	60.7	99.5	97.1
en→de	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	63.8	**100**	**100**
en→is	98.3	92.3	92.9	85.4	95.1	85.7	95.4	95.4	95.7	96.3	95.1	97.7	73.3	**98.9**	95.4
en→ru	97.4	99.0	99.5	99.5	98.4	98.4	98.4	99.5	99.0	**100**	98.4	99.5	53.9	**100**	97.4
en→zh	86.9	97.6	98.1	92.3	98.6	89.9	98.4	99.1	96.9	97.1	99	99.1	85.0	**99.4**	98.6
Average	96.5	97.1	97.3	94.8	97.9	94.2	98.0	98.2	97.9	98.3	97.8	99.2	67.3	**99.6**	97.7

Table 11: Mitigation rates MR in % (↑) for different post-hoc mitigation strategies on $\mathcal{D}_h^{dev}$ set.

	Fallback	MBR							Re-rank						
	NLLB	chrF	COMET				LaBSE		COMET			LaBSE			
	Beam	$t=1$	$t=1$	$t=1.5$	$t=1$	$t=2$	$t=1$	$t=1$	$t=1$	$t=1$	$t=0.8$	$t=1$	$t=1$	$t=1$	$t=0.8$
				$p=0.9$	$\epsilon=0.02$	$\epsilon=0.02$		$\epsilon=0.02$		$\epsilon=0.02$			MCBeam	$\epsilon=0.02$	$\epsilon=0.02$
en→cs	72.7	63.3	65.3	55	70.3	55.8	66.1	70.6	69.2	**73.5**	69.8	65.7	59.6	70	71.6
en→de	**76.8**	70.8	73.3	65.6	73.5	64.5	72.1	72.9	72.4	74.1	73.2	70.8	60.3	73.1	74.7
en→is	68.5	61.7	62.4	53.4	68.4	51.2	64.0	68.3	66.2	**71.2**	67.7	51.2	67.3	67.6	69.6
en→ru	71.4	65.1	67.7	57.7	72.4	56.2	68.2	72.4	70.1	**73.2**	70.8	66.8	57.6	71.0	72.9
en→zh	65.9	67.4	67.0	53.9	72.0	49.4	66.9	72.4	68.1	71.6	71.8	66.6	71.7	71.9	**74.0**
Average	71.1	66.9	67.1	57.1	71.3	55.4	67.5	71.3	69.2	**72.7**	70.7	64.2	63.3	70.7	72.6

Table 12: COMET scores (↑) for different post-hoc mitigation strategies on $\mathcal{D}_h^{dev}$ set.

5-19 of source sentences result in hallucinations when translating other target languages. A similar trend is observed across all language pairs. This indicates that the source sentences do not exhibit strong patterns that trigger hallucinations across different target languages.

	en→cs	*en→de*	*en→is*	*en→ru*	*en→zh*
en→cs	202	3	9	10	17
en→de	3	47	3	2	7
en→is	9	3	384	10	16
en→ru	10	2	10	186	17
en→zh	17	7	16	17	2178

Table 13: Number of common source sentences between $\mathcal{D}_h^{test}$ sets of different language pairs.

Translations In our analysis of hallucinated translations, we observed a substantial number of oscillatory hallucinations, characterized by repetitive sequences within the translation. These oscillatory hallucinations can be effectively identified using a top n-gram based hallucination detector Raunak et al., 2021, 2022; Guerreiro et al., 2023c,a. This detector flags a translation as a hallucination if the count of the top n-gram in the translation exceeds that of the source by a specified threshold. Based on prior works, we set n-gram to 4 and the threshold to 2. We find that 60% to 80% of the hallucinations were oscillatory in nature. The statistics for all language pairs are presented in Table 14.

en→cs	*en→de*	*en→is*	*en→ru*	*en→zh*
74.9%	76.9%	58.2%	60.7%	86.2%

Table 14: Oscillatory hallucination (%) in $\mathcal{D}_h^{test}$.

E Comparing Generation Methods for Post-hoc Mitigation strategies

Section 5.1 compares different mitigation strategies across different selection methods and utility metrics, focusing on the top performing sampling strategies. Here we compare different sampling strategies in Table 11 (MR) and Table 12 (COMET − wmt22-cometkiwi-da). Contrary to previous studies (Guerreiro et al., 2023c; Dale et al., 2023a) we find that MC-beam performs significantly worse than other sampling methods on both MR and COMET. We speculate that this is due to dropout not being used in the training of Llama-2, which is the backbone LLM for ALMA-7B-R. We find temperature $t = 1$ to perform best, with higher values of t significantly degrading both metrics. Using epsilon sampling with $\epsilon = 0.02$ consistently improves results.

F Common Hallucinations Before and After Fine-tuning

We compute the overlap in the hallucinated samples from ALMA-7B-R and $\mathcal{M}_{p+a}$ in Table 15. *Common source* column indicates the number of source sentences on which both baseline and fine-tuned mod-

	Hallucination Count		Count	
	ALMA-7B-R	$\mathcal{M}_{p+a}$	Common source	Common pairs (source+translation)
en→cs	179	4	2	0
en→de	39	1	0	0
en→is	388	35	10	4
en→ru	196	0	0	0
en→zh	2192	80	34	3

Table 15: Common source and (source, target) pairs between ALMA-7B-R and $\mathcal{M}_{p+a}$ on $\mathcal{D}_m^{test}$.

els hallucinate, while the *Common pairs* column reflects the number of identical (source, translation) pairs. For example, for *en→zh*, $\mathcal{M}_{p+a}$ generates 80 hallucinations on $\mathcal{D}_m^{test}$, of which 30 (37.5%) share the same source sentences that led to hallucinations in the baseline ALMA-7B-R. As expected, the percentage is lower when considering (source, translation) pairs, at 3.75%. It would be valuable to further investigate whether the high proportion of source sentences that still result in hallucinations after fine-tuning are due to underlying data quality issues, limitations in the modeling technique, or a combination of both.

G Hallucination Focused Preference Dataset Statistics

We report the character length statistics (mean, median, p95, and p99) for the source, preferred, and dispreferred samples in $\mathcal{D}_p^{train}$ in Table 16. Dispreferred samples have significantly longer lengths due to a large proportion of oscillatory hallucinations. Additionally, the hallucination score (HS) statistics (mean, median, p95, and p99) for the preferred and dispreferred data are shown in Table 17. We combine $\mathcal{D}_p^{train}$ with $\mathcal{D}_{alma}^{train}$ to fine-tune $\mathcal{M}_{p+a}$. Table 18 lists the dataset size of $\mathcal{D}_{alma}^{train}$.

H Statistics of Hallucination and Non-hallucination Samples

Table 19 shows source and translation character length statistics (mean, median, p95, and p99) for hallucination ($\mathcal{D}_h^{test}$) and non-hallucination ($\mathcal{D}_{nh}^{test}$) cases of the test set ($\mathcal{D}_m^{test}$), where translations are generated by ALMA-7B-R. We observe that the length statistics for source sentences are nearly identical between hallucination and non-hallucination samples. However, on the translation side, hallucinated translations are significantly longer than their non-hallucinated counterparts. For instance, the average length of hallucinated translations (839 characters) is 5.6 times longer than that of non-hallucinated translations (150 characters) across all language pairs. Additionally, for the non-hallucinated subset, the average source-to-target length ratio is nearly $1:1$, while for the hallucinated subset, it is $1:5.7$.

I Evaluation at Different Hallucination Score Thresholds

Our main evaluation results in Table 4 use a hallucination score threshold of 0.5. This threshold is also applied to create hallucination focused preference datasets. To assess whether our approach is biased toward this threshold, we re-evaluated both the baseline (ALMA-7B-R) and our best fine-tuned model ($\mathcal{M}_{p+a}$) at a few lower thresholds. It's important to note that as we lower the threshold, the distinction between hallucination and non-hallucination becomes increasingly blurred. However, a well-tuned model should still show improved performance over the baseline. Table 20 presents the evaluation results at different hallucination score thresholds (0.5, 0.45, and 0.4). While our $\mathcal{M}_{p+a}$ consistently outperforms ALMA-7B-R across all thresholds, the performance gap decreases as the threshold is lowered.

J Comparing Hallucination Score Improvement and Regression Rates Post Fine-tuning

Fine-tuning the LLM with our hallucination focused preference dataset significantly reduces the hallucination rate, but it doesn't provide insights into potential regressions in translation quality. For instance, the translation quality of individual samples might drop, but still remain within the threshold used to flag hallucinations. To explore this further, we plot a graph for the *en→zh* pair, as shown in Figure 2. The scatter plot represents each data point in our test set, where the X-axis denotes the hallucination score when translated by the ALMA-7B-R, and the Y-axis represents the score when the same sample is translated by $\mathcal{M}_{p+a}$. The diagonal line represents samples with no change in quality. Points above the line indicate a deterioration in translation quality post-fine-tuning, while points below the line indicate improvement. For *en→zh*, 28.4% of the samples showed a decline in quality after fine-tuning, whereas the majority (68.3%) improved. Similar trends were observed across all language pairs, as shown in the regression plots in Figure 5.

	Number of samples	Length											
		Mean			Median			p95			p99		
		x	y_p	y_d	x	y_p	y_d	x	y_p	y_d	x	y_p	y_d
en→cs	2063	168	190	1016	132	144	1102	434	511	1535	538	661	1972
en→de	671	156	199	1306	117	152	1258	426	554	2447	513	677	2770
en→is	3598	153	185	761	120	140	940	408	502	1245	549	677	1361
en→ru	1931	164	197	852	129	151	655	424	522	1522	543	673	1789
en→zh	8349	144	71	283	116	57	297	348	170	495	503	251	540
Average	3322	157	168	844	123	129	850	408	452	1449	529	588	1686

Table 16: Statistics of length in characters for source (x), preferred (y_p), and dispreferred (y_d) samples in $\mathcal{D}_p^{train}$.

	Hallucination Score							
	Mean		Median		p95		p99	
	y_p	y_d	y_p	y_d	y_p	y_d	y_p	y_d
en→cs	0.31	0.57	0.31	0.54	0.44	0.74	0.48	0.81
en→de	0.3	0.58	0.3	0.54	0.45	0.77	0.48	0.85
en→is	0.19	0.61	0.19	0.59	0.33	0.77	0.4	0.82
en→ru	0.22	0.65	0.22	0.65	0.35	0.81	0.41	0.84
en→zh	0.24	0.64	0.24	0.62	0.38	0.81	0.45	0.85
Average	0.25	0.61	0.25	0.59	0.39	0.78	0.44	0.83

Table 17: Statistics of hallucination score (HS) for preferred (y_p), and dispreferred (y_d) samples in $\mathcal{D}_p^{train}$.

en→cs	*en→de*	*en→is*	*en→ru*	*en→zh*
2009	2862	2009	2009	2783
cs→en	*de→en*	*is→en*	*ru→en*	*zh→en*
2009	2009	2009	2009	2009

Table 18: Number of samples in the preference dataset used by Xu et al. (2024) ($\mathcal{D}_{alma}^{train}$)

K Visualizing Hallucination and COMET Score Distributions

Distribution of scores Figure 3 and Figure 4 show the distribution of hallucination and COMET scores, respectively, for ALMA-7B-R and $\mathcal{M}_{p+a}$. We observe that the distribution of hallucination score for *en→{cs, is, zh}* shift slightly to the left after fine-tuning, indicating reduction in hallucination score. In contrast, the distributions for COMET are so closely overlapped that no definitive conclusions can be drawn.

Regression of scores Figure 5 and Figure 6 display regression plots for hallucination and COMET scores, respectively, comparing ALMA-7B-R and $\mathcal{M}_{p+a}$. The X-axis represents hallucination (or COMET) scores for ALMA-7B-R, while the Y-axis shows the scores for $\mathcal{M}_{p+a}$. The regression plots for hallucination clearly indicate improvements in the majority of translations across all language pairs, with the exception of *en→de*, which exhibits slightly higher regression. Conversely, the regression plots for COMET yield mixed results, making it challenging to draw definitive conclusions.

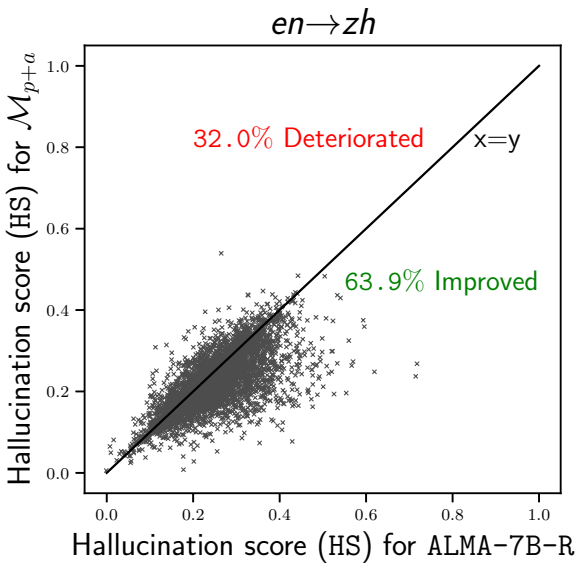

Figure 2: Regression plot showing translation quality on $\mathcal{D}_m^{test}$ for ALMA-7B-R and $\mathcal{M}_{p+a}$.

L Detailed General Translation Quality Evaluation

Section 5.2, Table 4 compares our fine-tuned models ($\mathcal{M}_p$ and $\mathcal{M}_{p+a}$) against ALMA-7B-R on WMT'23 *en→X* testsets using an average of three COMET models. In Tables 21, 22, 23, 24, we do a more detailed comparison, covering both *en→X* and *X→en* directions, WMT'22 and WMT'23 testsets and listing scores from individual COMET models as well as sacreBLEU.

	Source								Target							
	Mean		Median		p95		p99		Mean		Median		p95		p99	
	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$	$\mathcal{D}_{nh}^{test}$	$\mathcal{D}_{h}^{test}$
en→cs	148	186	127	153	334	453	479	532	157	1043	134	1090	355	1537	514	3583
en→is	148	147	127	118	334	380	479	537	173	751	144	933	389	1262	599	1505
en→ru	148	141	127	114	334	345	479	452	174	849	147	900	394	1512	572	1587
en→de	148	159	127	148	334	305	479	337	184	1274	157	1353	415	2127	596	2228
en→zh	148	144	127	116	334	349	479	510	63	280	50	297	146	476	315	520
Average	148	155	127	130	334	366	479	474	150	839	126	915	340	1383	519	1885

Table 19: Character length comparison for hallucination ($\mathcal{D}_{h}^{test}$) and non-hallucination subsets ($\mathcal{D}_{nh}^{test}$) of $\mathcal{D}_{m}^{test}$.

Threshold	0.5		0.45		0.4	
	ALMA-7B-R	$\mathcal{M}_{p+a}$	ALMA-7B-R	$\mathcal{M}_{p+a}$	ALMA-7B-R	$\mathcal{M}_{p+a}$
en→cs	179	4	380	45	1388	385
en→de	39	1	59	6	199	111
en→is	388	35	1271	353	4873	1722
en→ru	196	0	297	35	765	226
en→zh	2192	80	6024	608	17994	3967
Average count	599	24	1606	209	5044	1282
Average HR (%)	0.127	0.005	0.34	0.044	1.067	0.271

Table 20: Evaluation results at different HS threshold values: showing hallucination count and HR (%).

	BLEU	XCOMET	KIWI-22	KIWI-XXL	BLEU	XCOMET	KIWI-22	KIWI-XXL
	en→de				*en→cs*			
NLLB-3.3B	33.6	82.28	75.37	67.24	36.89	85.3	81.79	73.39
ALMA-7B-R	22.75	85.78	77.58	73.17	26.53	87.16	82.91	79.6
$\mathcal{M}_p$	23.04	84.7	77.65	71.7	28.91	86.66	82.43	76.89
$\mathcal{M}_{p+a}$	22.28	85.66	77.63	72.45	27.69	87.49	82.9	79.07
	en→ru				*en→zh*			
NLLB-3.3B	29.03	86.59	80.45	74.58	34.71	78.23	70.86	55.17
ALMA-7B-R	21.97	89.77	82.05	80.01	29.57	87.36	80.07	76.74
$\mathcal{M}_p$	22.99	87.94	81.49	77.49	34.09	87.41	80.03	75.36
$\mathcal{M}_{p+a}$	22.21	89.2	81.86	79.05	32.51	87.88	80.26	76.34
	en→X average							
NLLB-3.3B	**33.56**	83.1	77.12	67.59				
ALMA-7B-R	25.21	87.52	80.65	**77.38**				
$\mathcal{M}_p$	27.26	86.68	80.4	75.36				
$\mathcal{M}_{p+a}$	26.17	**87.56**	**80.66**	76.73				

Table 21: WMT'23 COMET and sacreBLEU scores for *en→X* directions. **XCOMET** = Unbabel/COMET-XCOMET-XXL, **KIWI-22** = Unbabel/COMET-wmt22-cometkiwi-da, **KIWI-XXL** = Unbabel/COMET-wmt23-cometkiwi-da-xxl. We reproduce all baseline model results. Best results per eval metric is shown in **bold** and second best is underlined.

	BLEU	XCOMET	KIWI-22	KIWI-XXL	BLEU	XCOMET	KIWI-22	KIWI-XXL
	de→en				*ru→en*			
NLLB-3.3B	35.26	81	77.69	72.96	31.74	84.17	79.88	77.1
ALMA-7B-R	28.59	84.71	78.68	76.08	31.78	88.94	80.97	80.57
$\mathcal{M}_p$	28.32	84.05	78.48	75.43	31.6	88.27	80.7	79.81
$\mathcal{M}_{p+a}$	28.31	85.01	78.66	76.06	31.69	88.67	80.94	80.35
	zh→en				*X→en* average			
NLLB-3.3B	22.15	82.77	77.15	71.89	**29.72**	82.65	78.24	73.98
ALMA-7B-R	22.51	89.01	79.6	77.63	27.63	87.55	**79.75**	**78.09**
$\mathcal{M}_p$	22.71	88.35	79.46	77.36	27.54	86.89	79.55	77.53
$\mathcal{M}_{p+a}$	22.5	88.99	79.57	77.79	27.5	**87.56**	79.72	78.07

Table 22: WMT'23 COMET and sacreBLEU scores for *X→en* directions. **XCOMET** = Unbabel/COMET-XCOMET-XXL, **KIWI-22** = Unbabel/COMET-wmt22-cometkiwi-da, **KIWI-XXL** = Unbabel/COMET-wmt23-cometkiwi-da-xxl. We reproduce all baseline model results. Best results per eval metric is shown in **bold** and second best is underlined.

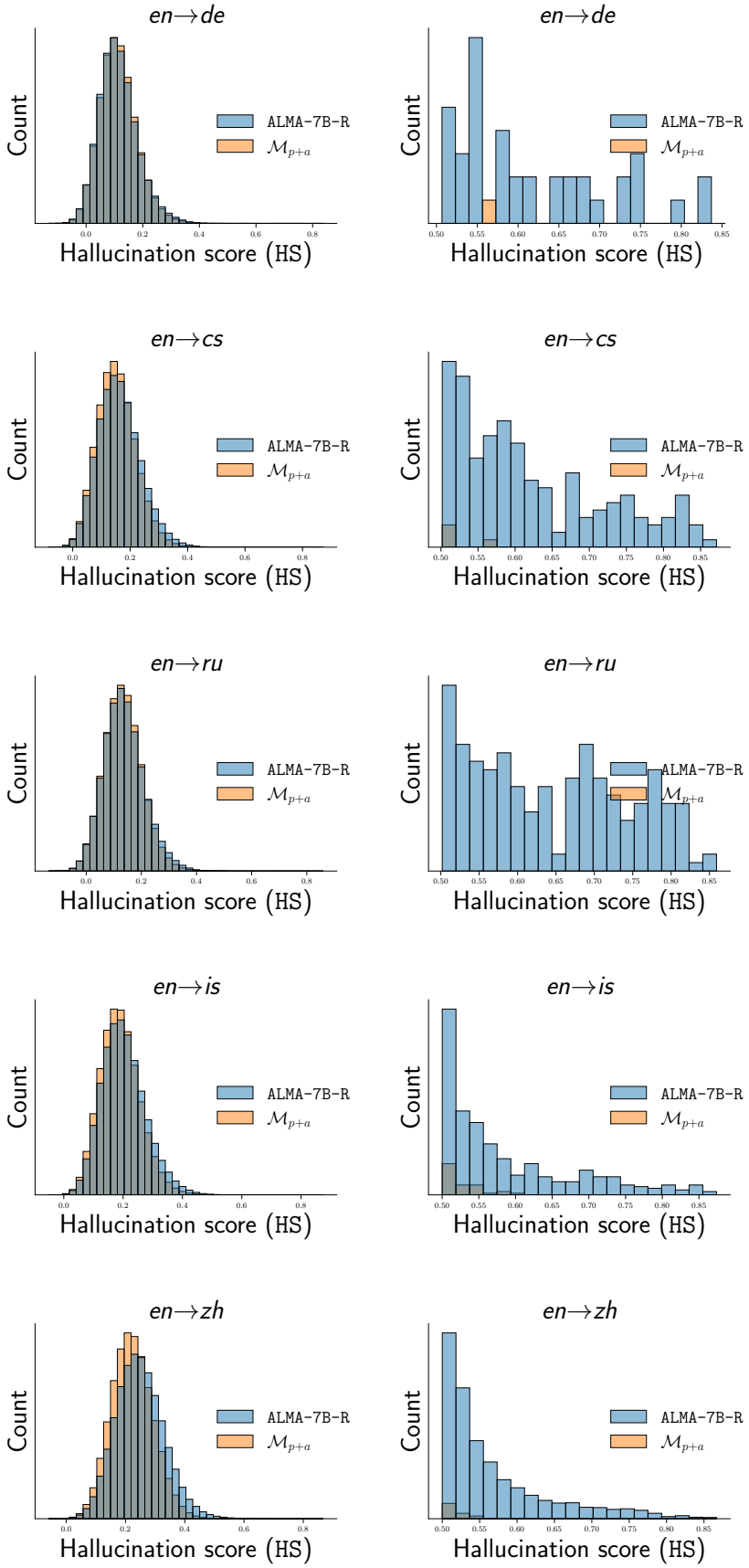

Figure 3: Hallucination score (HS) distribution for ALMA-7B-R and $\mathcal{M}_{p+a}$ on $\mathcal{D}_m^{test}$. Right plots are zoomed-in on hallucination regions.

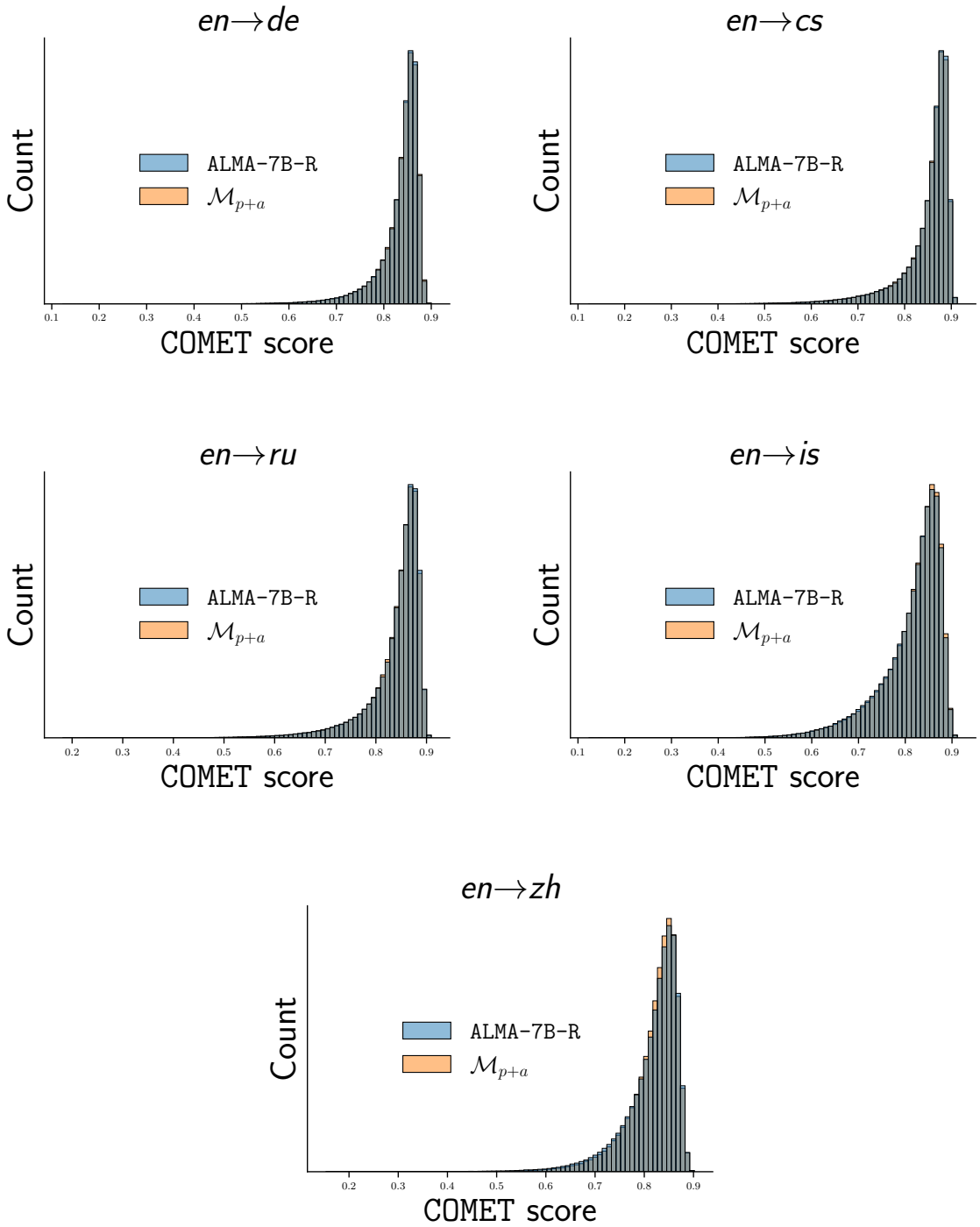

Figure 4: COMET score (`Unbabel/wmt22-cometkiwi-da`) distribution for `ALMA-7B-R` and $\mathcal{M}_{p+a}$ on $\mathcal{D}_m^{test}$.

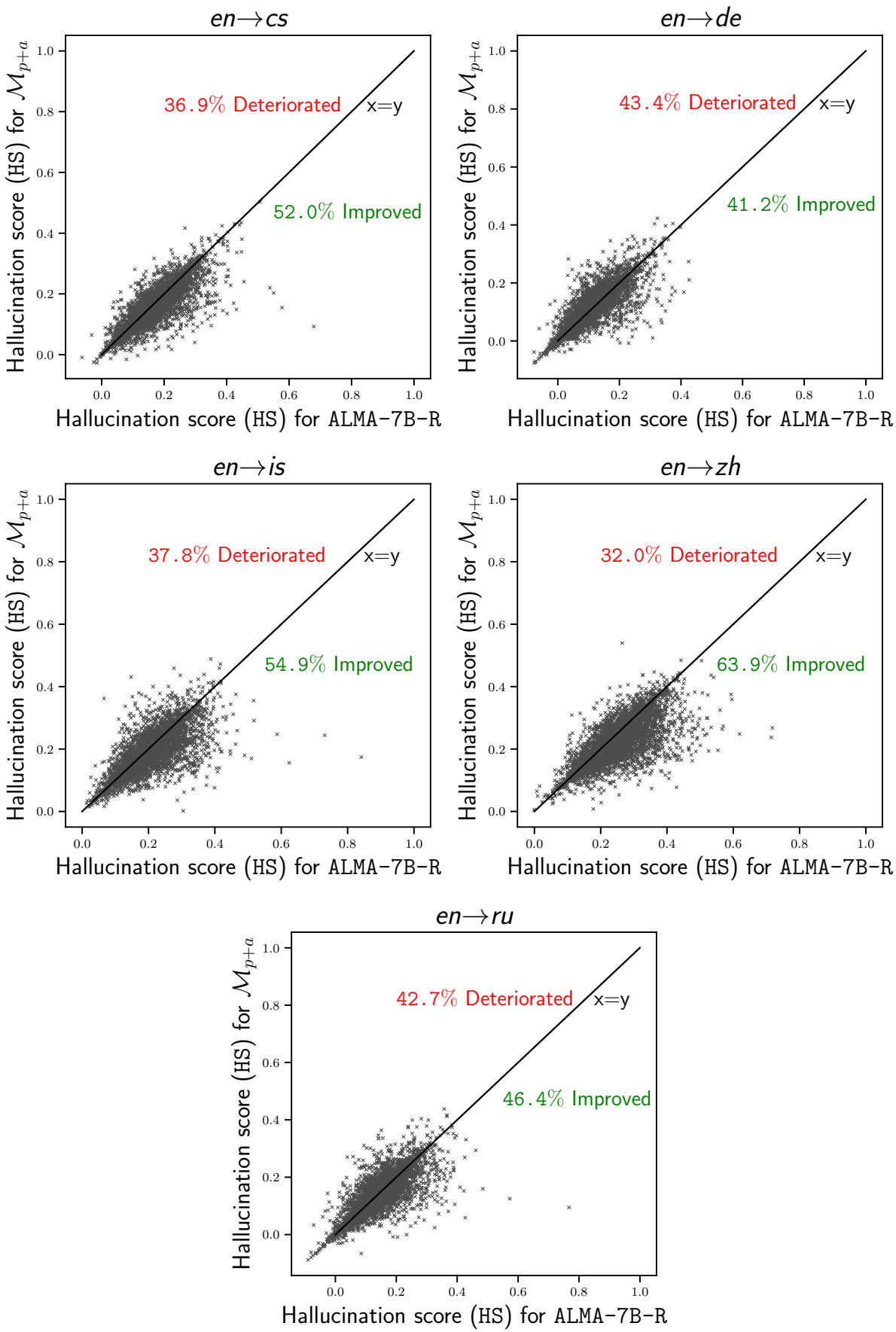

Figure 5: Regression plots showing hallucination score (HS) for ALMA-7B-R and $\mathcal{M}_{p+a}$ on $\mathcal{D}_m^{test}$.

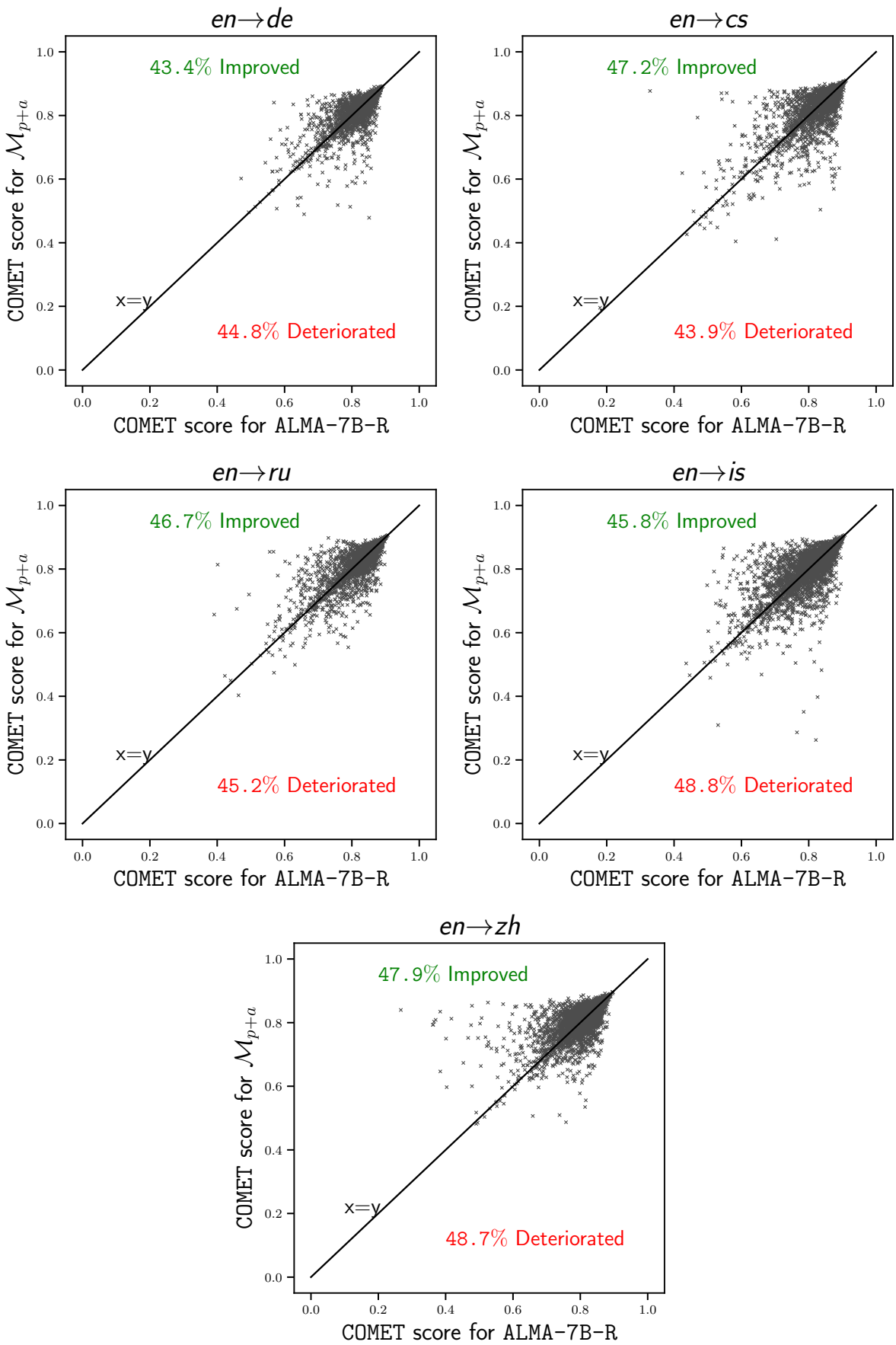

Figure 6: Regression plots showing COMET score (Unbabel/wmt22-cometkiwi-da) for ALMA-7B-R and $\mathcal{M}_{p+a}$ on $\mathcal{D}_m^{test}$.

	BLEU	XCOMET	KIWI-22	KIWI-XXL	BLEU	XCOMET	KIWI-22	KIWI-XXL
	en→de				*en→cs*			
NLLB-3.3B	34.16	95.62	83.35	82.36	36.27	89.29	84.15	81.65
ALMA-7B-R	27.01	96.68	83.41	83.94	25.21	90.24	84.95	86.49
$\mathcal{M}_p$	27.7	95.84	83.26	82.58	27.26	89.67	84.45	83.87
$\mathcal{M}_{p+a}$	27.52	96.4	83.21	83.24	25.65	90.37	84.8	85.6
	en→is				*en→zh*			
NLLB-3.3B	23.46	79.3	79.63	75.42	31.91	81.42	75.05	65.62
ALMA-7B-R	20.81	85.45	81.53	83.94	30.5	89.66	81.88	82.77
$\mathcal{M}_p$	22.19	86.71	81.74	83.33	32.57	89.87	81.9	81.57
$\mathcal{M}_{p+a}$	22.11	87.26	81.68	83.71	31.85	90.31	82.08	82.67
	en→ru				*en→X* average			
NLLB-3.3B	30.22	91.08	83.35	82.35	31.2	87.34	81.11	77.48
ALMA-7B-R	23.43	93.35	84.04	86.5	25.39	<u>91.08</u>	<u>83.16</u>	**84.72**
$\mathcal{M}_p$	24.93	92.3	83.8	84.45	**26.93**	90.88	83.03	83.16
$\mathcal{M}_{p+a}$	23.79	93.19	84.06	86.15	<u>26.18</u>	**91.51**	**83.17**	<u>84.27</u>

Table 23: WMT'22 COMET and sacreBLEU scores for *en→X* directions. **XCOMET** = Unbabel/COMET-XCOMET-XXL, **KIWI-22** = Unbabel/COMET-wmt22-cometkiwi-da, **KIWI-XXL** = Unbabel/COMET-wmt23-cometkiwi-da-xxl. We reproduce all baseline model results. Best results per eval metric is shown in **bold** and second best is <u>underlined</u>.

	BLEU	XCOMET	KIWI-22	KIWI-XXL	BLEU	XCOMET	KIWI-22	KIWI-XXL
	de→en				*cs→en*			
NLLB-3.3B	29.45	91.35	81.02	82.11	49.03	85.94	81.72	80.25
ALMA-7B-R	31.32	93.6	81.4	83.61	43.71	89.32	82.37	82.91
$\mathcal{M}_p$	31.09	93.2	81.18	82.76	43.44	88.88	82.19	81.96
$\mathcal{M}_{p+a}$	30.94	93.69	81.31	83.31	42.94	89.6	82.36	82.57
	is→en				*zh→en*			
NLLB-3.3B	34.27	74.8	79.87	79.22	20.96	82.28	75.38	68.36
ALMA-7B-R	38.86	86.6	81.49	85.63	22.32	89.47	78.9	76.5
$\mathcal{M}_p$	39.32	86.43	81.42	85.63	22.1	88.79	78.61	75.95
$\mathcal{M}_{p+a}$	38.61	86.54	81.41	85.54	22.08	89.25	78.68	76.31
	ru→en				*X→en* average			
NLLB-3.3B	40.17	89.43	80.87	78.39	34.78	84.76	79.77	77.67
ALMA-7B-R	38.91	92.27	81.57	81.22	**35.02**	<u>90.25</u>	**81.15**	**81.97**
$\mathcal{M}_p$	39.1	91.94	81.35	80.8	<u>35.01</u>	89.85	80.95	81.42
$\mathcal{M}_{p+a}$	38.47	92.54	81.55	81.08	34.61	**90.33**	<u>81.06</u>	<u>81.76</u>

Table 24: WMT'22 COMET and sacreBLEU scores for *X→en* directions. **XCOMET** = Unbabel/COMET-XCOMET-XXL, **KIWI-22** = Unbabel/COMET-wmt22-cometkiwi-da, **KIWI-XXL** = Unbabel/COMET-wmt23-cometkiwi-da-xxl. We reproduce all baseline model results. Best results per eval metric is shown in **bold** and second best is <u>underlined</u>.