Paying More Attention to Source Context: Mitigating Unfaithful Translations from Large Language Model

Hongbin Zhang[†‡]**,Kehai Chen**[†*]**,Xuefeng Bai**[†]**,Yang Xiang**[‡]**,Min Zhang**[†]

[†]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[‡]Peng Cheng Laboratory, Shenzhen, China
azure.starzhang@gmail.com,{chenkehai,baixuefeng,zhangmin2021}@hit.edu.cn,
xiangy@pcl.ac.cn

Abstract

Large language models (LLMs) have showcased impressive multilingual machine translation ability. However, unlike encoder-decoder style models, decoder-only LLMs lack an explicit alignment between source and target contexts. Analyzing contribution scores during generation processes revealed that LLMs can be biased towards previously generated tokens over corresponding source tokens, leading to unfaithful translations. To address this issue, we propose to encourage LLMs to pay more attention to the source context from both source and target perspectives in zeroshot prompting: 1) adjust source context attention weights; 2) suppress irrelevant target prefix influence; Additionally, we propose 3) avoiding overreliance on the target prefix in instruction tuning. Experimental results from both human-collected unfaithfulness test sets focusing on LLM-generated unfaithful translations and general test sets, verify our methods' effectiveness across multiple language pairs. Further human evaluation shows our method's efficacy in reducing hallucinatory translations and facilitating faithful translation generation. [1]

1 Introduction

Large language models (LLMs; Brown et al. 2020; Liu et al. 2023) have shown great potential in machine translation within recent years (Lin et al., 2022; Zhang et al., 2022; Hendy et al., 2023; Jiao et al., 2023b). Given the different modeling architectures and pre-trained objectives in decoder-only LLMs and encoder-decoder neural machine translation models, previous studies have probed into leveraging LLMs for translation via in-context learning (Zhu et al., 2023; Vilar et al., 2023; Zhang et al., 2023a) or instruction tuning (Jiao et al.,
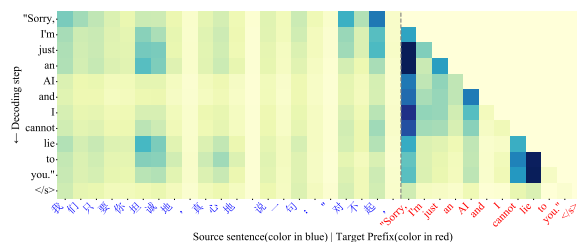

Figure 1: Contribution visualization of a Zh⇒En unfaithful translation instance. Each predicted token (row) corresponds to the contribution of each input token including source tokens and target prefixes (column) to the output token. One of the correct translations of the given source sentence is "*We just want you to honestly and sincerely say Sorry*".

2023a; Muennighoff et al., 2023; Xu et al., 2023; Alves et al., 2024).

However, decoder-only LLMs, lacking an explicit mechanism, such as cross-attention modules (Bahdanau et al., 2015; Vaswani et al., 2017) in encoder-decoder architectures, for aligning the source and target context. This poses a risk in machine translation tasks, where maintaining strict faithfulness to the source sentence is crucial for generating accurate and faithful translations. In Figure 1, we visualize the influence of the source and target tokens on the generating tokens during the generation process using contribution scores(developed from Ferrando et al. (2022a,b) and be detailed in Apendix A) in decoder-only LLMs, e.g., Llama-2-7B. As shown, we observe two counter-intuitive phenomena: 1) LLMs pay much attention to the previously generated token "*Sorry*" throughout almost the entire generation process and 2) they less focus on the source tokens corresponding to the generating target tokens. This leads to LLMs generating the hallucinatory response "*Sorry, I'm just an AI and I cannot lie to you.*", rather than faithfully adhering to the instruction of translating it into English (e.g., "*We just want you to honestly and sincerely say Sorry*").

13816

To tackle this issue, we propose strategies targeting both source and target aspects to guide the decoder-only LLMs toward focusing more on the source context during the generation process. Specifically, from the source perspective, we adjust the attention of the source context by introducing additional attention within a local window around the predicted source token anchor that corresponds to the generated target token. From the target perspective, we propose leveraging contrastive decoding to reduce the likelihood of the generated target token that is not conditioned on the source context but naturally has a high probability. Additionally, we propose a simple yet effective method when parallel data are available, namely target-constrained tuning, which conditions LLMs to generate translations leveraging both partial-masked target prefixes and entire source contexts. Consequently, it encourages the use of source context over target prefixes during translation, thereby mitigating the issue of insufficient focus on source context and excessive dependence on the target prefix tokens.

We take LLaMA-2 series (Touvron et al., 2023) as backbones and conduct experiments in both unfaithful translation test sets and open benchmarks, like WMT22 (Kocmi et al., 2022) and Flores (Goyal et al., 2022). Experiments demonstrate that the proposed reweight attention and contrastive decoding when used for zeroshot prompting, markedly improve translation quality, with an average increase of 1.7 BLEU and 4.0 COMET scores compared to vanilla prompting. Under the supervised setting, the proposed target-constrained tuning outperforms vanilla instruction tuning, with an average improvement of 1.1 in BLEU score and 0.6 in COMET score. Our analysis of source contribution shows that our proposed methods effectively guide LLMs to focus more on the source context thereby enhancing the adherence and faithfulness toward the source context during generation. Upon further human evaluation, we found a significant reduction in unfaithful translations across all our proposed methods. Our main contributions are summarized as follows:

- This paper first focuses on the issue that the LLM-based MT over-depends on the generated target-side contextual information due to lacking cross-attention, which leads to more unfaithful translation.

- This paper proposes three methods aimed at different application scenarios to improve this serious phenomenon of unfaithful translation brought by the target-side context bias.

- We annotate a specific unfaithful dataset tailored for LLMs to evaluate the effectiveness of the proposed approach.

2 Methodology

Recognizing the significance of both source and target aspects in machine translation, we address the aforementioned issues by enhancing the contribution of the source context and diminishing the influence of the target prefixes. Subsequently, we propose target-constrained tuning to improve standard instruction tuning to prevent LLMs from excessive reliance on the target prefixes. The overview of our proposed methods is shown in Figure 2.

2.1 Boosting Source Influence: Reweight Attention

Our reweight attention mechanism is based on a local window, drawing inspiration from Luong et al. (2015). This mechanism selectively focuses on the subset of the source context during translation. In more detail, the model initially determines an aligned position p_t for each target token at time t. The local attention weight is subsequently derived from query vectors and key vectors corresponding to the source context within the window $[p_t - D, p_t + D]$, where the value of D is empirically determined. Subsequently, we explore two variants of the method, outlined below:

Monotonic Alignment: We straightforwardly set $p_t = t$, assuming a rough monotonic alignment between source and target sequences.

Contribution-guided Alignment: We propose leveraging the contribution measurement introduced by Kobayashi et al. (2021) to heuristically designate the most significant source token from the entire source context as the aligned position.

$$p_t = \max_i ||(\text{LN}(\sum_{i=1}^{S} T_t(x_i) + b_O + x_t))||_2, \quad (1)$$

Here, $T_t(x_i)$ represents the linear transformation detailed in Appendix A, LN represents the layer normalization operation, b_O represents the bias term of the output linear projection, x_t represents

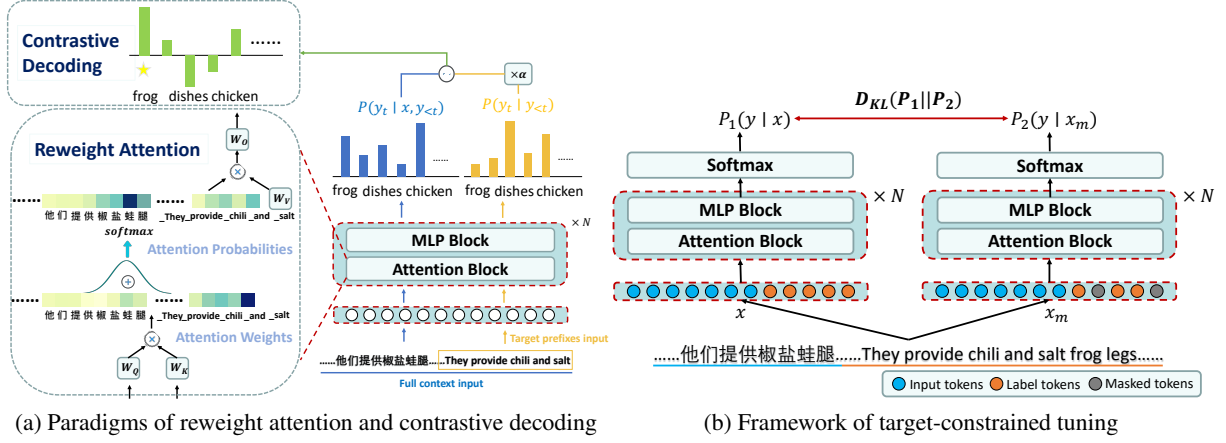(a) Paradigms of reweight attention and contrastive decoding   (b) Framework of target-constrained tuning

Figure 2: The left picture shows the paradigms of proposed unsupervised methods, including the reweight attention and contrastive decoding. The right picture illustrates the target-constrained tuning, detailing how the two different inputs, full input x and label-masked input x_m will go through the model and obtain two distributions P_1 and P_2.

the residual connection and S denotes the length of the source sentence.

To promote alignment points near p_t, we model a Gaussian distribution centered around p_t. Specifically, the alignment vectors α_t are defined as:

$$\alpha_t(s) = \omega \times exp(-\frac{(s - p_t)^2}{2\sigma^2}), \quad (2)$$

Here, ω serves as the scale factor regulating the additional attention weight, s represents the index of the source tokens, and the standard deviation σ is empirically set to $\frac{D}{2}$. Subsequently, we modify the attention output by adding extra attention weights calculated by the local attention window:

$$\text{attention}(Q_t, K_x, V_x) = \text{softmax}((1 + \alpha_t) \times \frac{Q_t K_x^T}{\sqrt{d_k}})V_x, \quad (3)$$

where $Q_t = W_Q x_t, K_x = W_K x, V_x = W_V x$ denote the query vector of target position t, key vectors, and value vectors of input x, respectively. d_k represents the dimension of a single vector.

2.2 Mitigating Target Impact: Contrastive Decoding

To prevent undesired generations that are not conditioned on the source context, we propose facilitating LLMs to diminish the contribution of the target prefix through contrastive decoding (Li et al., 2016; Shi et al., 2023).

By replacing the standard log-likelihood objective function with the maximum mutual information (MMI) as an alternative objective function O, we select tokens that maximize the mutual information between the input context X and the

translation output Y:

$$O_{MMI} = \log \frac{P(X, Y)}{P(X)P(Y)}, \quad (4)$$

$P(\cdot)$ is estimated by providing the LLM with the translation instruction prompt as shown in Appendix B. This prevents bias towards translations that may inherently carry a high probability without being conditioned on the source context. Instead, it encourages responses that are specifically tailored to the given source input. Moreover, we extend the MMI objective which introduces a hyperparameter α that controls the degree to which unconditional responses are penalized:

$$y_t = \arg\max_{v}\{\log p(y_t \mid x, y_{<t}) - \alpha \log p(y_t \mid y_{<t})\}, \quad (5)$$

where x is the input query, and $y_{<t}$ is the response before timestep t.

2.3 Target-constrained Instruction Tuning

Based on the previous analysis and inspired by Bengio et al. (2015) and Liang et al. (2021), we propose target-constrained instruction tuning to encourage LLMs to learn generating translations given the entire source context and incomplete target prefixes, thereby preventing over-relying on target prefixes when generation.

Concretely, given the instruction style query x which contains the translation instruction and source sentence as constructed in Appendix B, and target sentence y as the label for supervised training, we first feed the full instruction $\{x, y\}$ to go through the forward pass of the model to obtain the distribution of the model predictions

denoted as $\mathcal{P}^f(y_t|x, y_{<t})$. We then generate the partially masked targets y^m, where target tokens are masked with a probability β. Following this, the target-constrained instruction input $\{x, y^m\}$ is fed into the model, resulting in a target-constrained distribution for the model's prediction, represented as $\mathcal{P}^c(y_t|x, y_{<t}^m)$. During the training step, our method aims to regularize model predictions by minimizing the bidirectional Kullback-Leibler(KL) divergence between the output distributions corresponding to the same source context, which is:

$$\mathcal{L}_{KL} = \frac{1}{2}\left(\mathcal{D}_{KL}\left(\mathcal{P}^f\left(y_t \mid x, y_{<t}\right) \| \mathcal{P}^c\left(y_t \mid x, y_{<t}^m\right)\right) + \mathcal{D}_{KL}\left(\mathcal{P}^c\left(y_t \mid x, y_{<t}^m\right) \| \mathcal{P}^f\left(y_t \mid x, y_{<t}\right)\right)\right).$$
(6)

Building upon the basic negative log-likelihood learning objective $\mathcal{L}_{NLL}$ associated with the two forward passes:

$$\mathcal{L}_{NLL} = -\log\mathcal{P}^f\left(y_t \mid x, y_{<t}\right) - \log\mathcal{P}^c\left(y_t \mid x, y_{<t}^m\right),$$
(7)

To sum up, we jointly optimize the total loss, incorporating full context translation loss, target-constrained translation loss, and regularized KL-Divergence loss, as illustrated below:

$$\mathcal{L} = \mathcal{L}_{NLL} + \lambda \cdot \mathcal{L}_{KL},$$
(8)

where λ is the coefficient weight to control $\mathcal{L}_{KL}$. By minimizing this loss, the probability distribution of the entire input context becomes less dependent on the target prefixes, thereby encouraging LLMs to utilize the source context to the greatest extent for generating translation.

3 Experiments

We conduct experiments on the proposed human-collected unfaithful translation test sets containing unfaithful translations covering three languages and four translation directions. Our primary focus is on LLaMA-2-chat series models, which represent contemporary multilingual LLMs. More details of experimental settings can be found in Appendix B. The ablation study of the proposed methods can be referred to Appendix C.

3.1 Experimental Settings

Dataset We heuristically gather translation data that is prone to be unfaithful or hallucinatory based on the metric detailed in the Appendix A for all four translation directions(Chinese⇔English and German⇔English) as the evaluation data.

We utilize human-written data from past WMT competitions rather than public training data to prevent the introduction of noises into instruction tuning following Jiao et al. (2023a) and Xu et al. (2023). We employ newstest2017-2021 of Chinese⇔English and newstest2014-2021 of German⇔English tasks (Post, 2018), This yields a total of 62.9K training sentence pairs data for all four directions.

Baseline We train the model in a bilingual translation manner separately for different translation directions and use LLaMA-2-7B-chat as our backbone model given its best zero-shot and instruction following performance.

Vanilla Instruction Tuning/Vanilla Instruction Tuning LoRA *Full-Weight* or *LoRA* vanilla instruction tuning on high-quality parallel data for LLaMA-2-7B-chat.

Scheduled Sampling Tuning LoRA/R-Drop Tuning LoRA *LoRA* instruction tuning using Schedule Sampling (Bengio et al., 2015) or R-Drop (Liang et al., 2021) on high-quality parallel data for LLaMA-2-7B-chat.

Metics. We use BLEU (Papineni et al., 2002) implemented in SacreBLEU[2] (Post, 2018), and COMET[3] (Rei et al., 2020) from *Unbabel/wmt22-comet-da*[4] for automatic evaluation.

3.2 Main Results

The results in Table 1 reveal that the zeroshot prompting of the LLaMA2-7b-chat model exhibits poor performance on the unfaithful translation dataset. Comparing the baseline results with the improved outcomes achieved by our proposed methods across various translation directions and languages, we noted the following:

Elevating source focus brought improved translation quality. The reweight attention method outperforms vanilla zeroshot prompting, showing an average improvement of 2.1 BLEU and 4.7 COMET. It also exhibits superior performance in translations from English compared to translations to English. This observation might be attributed to the more severe inability to pay sufficient attention to specific source sentences when translating to languages other than English, leading to a decline in translation performance. The results suggest that

System	De ⇒ En		En ⇒ De		Zh ⇒ En		En ⇒ Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Unsupervised Setting								
Vanilla Zeroshot	23.2	77.8	7.11	60.6	11.4	74.3	3.81	42.8
Reweight Attention (RA)	**24.5**	**79.0**	**10.6**	**63.3**	**12.5**	**75.0**	**6.14**	**57.0**
Contrastive Decoding (CD)	24.2	78.7	9.10	61.5	12.1	74.6	4.81	53.7
Supervised Setting								
Vanilla Fewshot	27.1	81.7	15.7	74.2	14.2	76.5	16.2	73.1
Vanilla Instruction Tuning	27.3	82.2	19.8	77.5	15.7	76.4	17.5	74.8
Target-constrained tuning	29.6	83.0	20.9	78.0	16.1	77.0	17.9	75.2
Vanilla Instruction tuning LoRA	29.1	82.9	20.3	78.5	15.5	76.5	18.6	76.2
Scheduled Sampling tuning LoRA	30.5	83.1	20.2	78.5	16.5	76.8	18.6	76.1
R-Drop tuning LoRA	30.3	83.1	20.1	**78.8**	**16.6**	**77.1**	18.7	**76.5**
Target-constrained tuning LoRA	**30.8**	**83.2**	**20.6**	78.5	**16.6**	77.0	**19.1**	**76.5**

Table 1: Translation performance of LLaMA2-7b-chat model on human-collected unfaithful translation test sets. The bold number marks the best metric results from the methods under the same translation evaluation setting.

the proposed reweight attention can enhance translation quality by directing LLaMA to prioritize the aligned source context.

Mitigation of target influence generates better translation. The contrastive decoding strategy significantly improves the translation performance of LLMs, outperforming the baseline with an average improvement of 1.2 BLEU and 3.3 COMET. However, the extent of improvement varies across different translation directions. It is noteworthy that contrastive decoding markedly enhances translations between German and English, while it results in only a marginal improvement in translations between Chinese and English compared to vanilla prompting.

Target-constrained tuning refines vanilla instruction tuning. Instruction tuning only slightly improves translation performance compared to fewshot prompting, suggesting its limited effectiveness in addressing insufficient focus on the source context. However, the proposed target-constrained tuning consistently outperforms vanilla instruction tuning, with an average gain of 1.05 BLEU and 0.58 COMET. Furthermore, we employ low-rank adaptation (Hu et al., 2022) to fine-tune the partial parameters of LLMs, aiming for improved efficiency. Experimental results demonstrate that LoRA tuning enhances performance across all translation directions and outperforms all parameters tuning. This is likely attributed to the limited tunable parameters in LoRA, preventing LLMs from overfitting to the small translation dataset and enhancing generalization ability (Jiao et al., 2023a). The results of the two baselines show
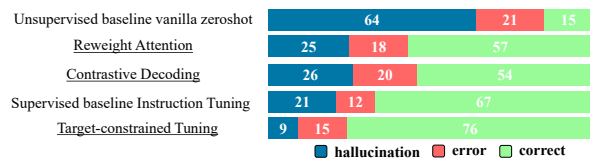

Figure 3: Human annotation results: percentages of translation categories for different methods.

that while both the Scheduled Sampling (Bengio et al., 2015) and R-drop (Liang et al., 2021) methods can enhance performance to some degree, they are not as effective in mitigating unfaithful translations as our proposed target-constrained method.

3.3 Human Evaluation

Despite the utility of automatic evaluation metrics, they do not explicitly measure the degree to which our proposed method has mitigated the existence of unfaithful or hallucinated spans. Therefore, we conducted a human evaluation. Moreover, the consistency of hallucination evaluation between human evaluation and automatic metrics is studied in Appendix E.

Data. Initially, we collected 100 instances, selected from the instances exhibiting the lowest cross-lingual sentence similarity between hypotheses and source sentences (Dale et al., 2023; Feng et al., 2022). These sentences are likely to have hallucinated spans in the unfaithful translation dataset and were used as the human evaluation set. In total, 64% of the sentences had their original translations marked as translation hallucination, 21% as translation error, and 15% as correct translations. Next, we use each method to translate

the same set of 100 source sentences. The resultant sentence pairs are categorized by three annotators into three groups: Correct, Error, and Hallucination. More details on the annotation guidelines and inter-annotator agreement can be found in Appendix D.

Results. The human evaluation results are displayed in Figure 3. All proposed methods decrease the translation hallucination rate by at least a factor of 2.5. The unsupervised methods, including the reweight attention and contrastive decoding, significantly reduce hallucination in translations compared to baseline vanilla zeroshot. Interestingly, in terms of mitigating translation hallucinations, both the unsupervised methods perform on par with supervised instruction tuning. However, the baseline instruction tuning exhibits fewer errors, which is anticipated, given that it was trained on a parallel corpus to generate accurate translations. When it comes to target-constrained tuning, we observe that it produces significantly fewer translation hallucinations than vanilla instruction tuning, but it results in more erroneous translations. Our analysis of the annotations reveals that while target-constrained tuning is capable of generating source-related translations, it also introduces some ambiguity issues, resulting in more errors. Overall, the application of all our proposed methods significantly reduces translation hallucinations.

3.4 Generalize to other settings

System	En ⇒ De		Zh ⇒ En	
Base Model:LLaMA2-7b-Chat	BLEU	COMET	BLEU	COMET
Flores 101 Test Set				
Vanilla Zeroshot	21.8	78.7	19.3	83.9
Reweight Attention	21.9	78.7	19.8	84.0
Contrastive Decoding	21.7	78.5	19.5	84.1
Vanilla Instruction tuning LoRA	29.2	84.6	22.9	84.6
Target-constrained tuning LoRA	29.5	84.7	24.5	85.1
WMT 22 Test Set				
Vanilla Zeroshot	23.5	74.4	19.7	77.4
Reweight Attention	23.7	74.5	20.1	77.4
Contrastive Decoding	23.4	74.4	19.7	77.5
Vanilla Instruction tuning LoRA	34.0	81.7	24.4	77.6
Target-constrained tuning LoRA	35.3	82.1	25.3	78.5

Table 2: Translation performance of LLaMA2-7b-chat model on Flores101 and WMT22 test sets

On general open dataset. While we have previously tested our methods on the proposed unfaithful translation dataset, we aim to determine their effectiveness on general open MT benchmarks. These include test sets such as Flores (Goyal et al., 2022) and WMT22 (Kocmi et al., 2022). As

System	En ⇒ De		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET
Base Model:BLOOMZ-7b1-mt				
Vanilla Instruction tuning LoRA	20.5	64.2	25.5	79.1
Target-constrained tuning LoRA	21.2	64.4	26.2	79.4
Base Model:ChatGLM3-6b				
Vanilla Instruction tuning LoRA	27.1	73.2	26.8	79.1
Target-constrained tuning LoRA	27.2	73.2	27.2	79.2
Base Model:Vicuna-7b				
Vanilla Instruction tuning LoRA	32.5	80.6	25.0	76.8
Target-constrained tuning LoRA	33.5	81.3	25.2	76.9

Table 3: Translation performance of various families of LLMs with a similar size on WMT22 test sets.

System	En ⇒ De		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET
Base Model:LLaMA2-7b-chat				
Vanilla Instruction tuning LoRA	34.0	81.7	25.5	78.3
Target-constrained tuning LoRA	34.5	81.8	25.8	78.4
Base Model:LLaMA2-13b-chat				
Vanilla Instruction tuning LoRA	37.0	83.2	27.6	79.3
Target-constrained tuning LoRA	37.6	83.4	27.8	79.4
Base Model:LLaMA2-70b-chat				
Vanilla Instruction tuning LoRA	41.3	84.8	30.2	80.5
Target-constrained tuning LoRA	41.8	85.0	31.1	80.8

Table 4: Translation performance of various families of LLMs with a similar size on WMT22 test sets.

shown in Table 2, our proposed methods including the reweight attention and contrastive decoding still yield results comparable to vanilla zeroshot prompting. Target-constrained tuning continues to be effective in enhancing performance, given its dual function as a strategy to prevent overfitting on smaller translation datasets. Although there is an imbalanced improvement between the unfaithful test set and general test sets, our proposed methods can still achieve comparable performance on less hallucinated translation test sets. For a more in-depth analysis of this phenomenon, please refer to Appendix F.

Employing LLMs with different families and scales. We subsequently apply our target-constrained tuning to various LLM families that are of a similar size to LLaMA2-7b and evaluated on the WMT22 test sets, as shown in Table 3. Despite the variations in architecture and pretraining corpus among these models, our target-constrained tuning method proves to be universally effective. In considering the different scales of LLMs, we select various scales of LLaMA2-chat versions to compare their performance on the WMT22 test sets, as depicted in Table 4. Our proposed target-constrained tuning notably outperforms the vanilla instruction tuning. We have also conducted additional experiments on the unfaithful test set, with details provided in Appendix H.

(a) Contribution of the source context at each generation step
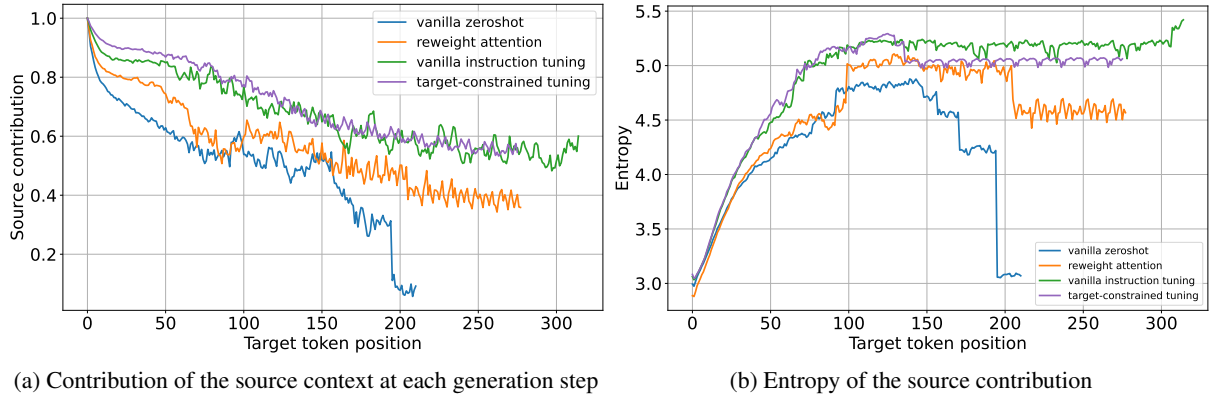
(b) Entropy of the source contribution

Figure 4: For each generation step, the figure shows the (a) contribution and (b) entropy of source context in the translation direction from Chinese to English. The points on the lines denotes the average score across the i-th target token position. Note that different methods result in different target generation lengths.

4 Analysis

4.1 Contribution Distribution

In this section, we reveal the contribution characteristic of input tokens' to generation observed in the behavior of the LLaMA2-7b-chat model. Specifically, during the translation from Chinese to English using our proposed unfaithful dataset, we compute the average contribution of source context generation and the entropy of the source contribution from each target token position.

Changes of source contribution during generation. Following section A, we can derive the input token contribution matrix $C(t, i)$, which denotes the contribution score of the i_{th} input token during the generation of the t_{th} target token. Specifically, for each generation step t, we compute the aggregate contribution from the source as $C_t(source) = \sum_{i \in S} C(t, i)$, where S is the set of the indices corresponding to source tokens. As illustrated in Figure 4a, we note that, during the entire generation process, the impact of the source diminishes (or, conversely, the impact of the prefix intensifies). This is an anticipated outcome: as the prefix lengthens, the model faces less uncertainty in determining which source tokens to utilize but needs to exert more control over fluency. These observations align with findings from previous work by Voita et al. (2021). It is evident that during the early phases of token generation, the source contributions of the reweight attention method surpass those of the vanilla zeroshot, and the source contributions of the target-constrained tuning exceed those of the baseline instruction tuning. Furthermore, there's a steep decline

observed in the baseline zeroshot, further detailed in Appendix G. In contrast, our methodologies show a gradual decline. This suggests that our proposed methodologies can effectively amplify the significance of the source, and it also indicates the effectiveness of our methods in mitigating the issue associated with inadequate attention to the source context.

Entropy of source contributions. Let's examine the 'sharpness' of the contributions of source tokens at different steps of generation. For each step, denoted by t, we calculate the entropy of the normalized source contributions, represented by $\{\frac{C(t, i)}{C_t(source)}\}_{i=1}^{S}$. As shown in Figure 4b, during the zeroshot generation phase of LLMs, we observe an increase in entropy until about the 150th position in the generated translation. Beyond this point, the entropy starts to diminish as the rest of the translation is generated. However, the baseline instruction tuning as well as our proposed methods invert this trend. After the generation of about the 150-th token, the entropy of source contributions remains high and fluctuates around 5. This indicates that when confronted with a longer context, the generation of LLMs necessitates a broader source context. Furthermore, our proposed methods increase the entropy of source contributions compared to the vanilla methods during the initial stage of translation generation. This implies that the application of our methods in generating translations requires a more comprehensive view of the source context, thereby enhancing the significance of the contribution from the source context.

(a) Vanilla zero-shot prompting over-translation case.

(b) Reweight attention mitigate translation hallucination issue

(c) Vanilla instruction tuning under-generation case

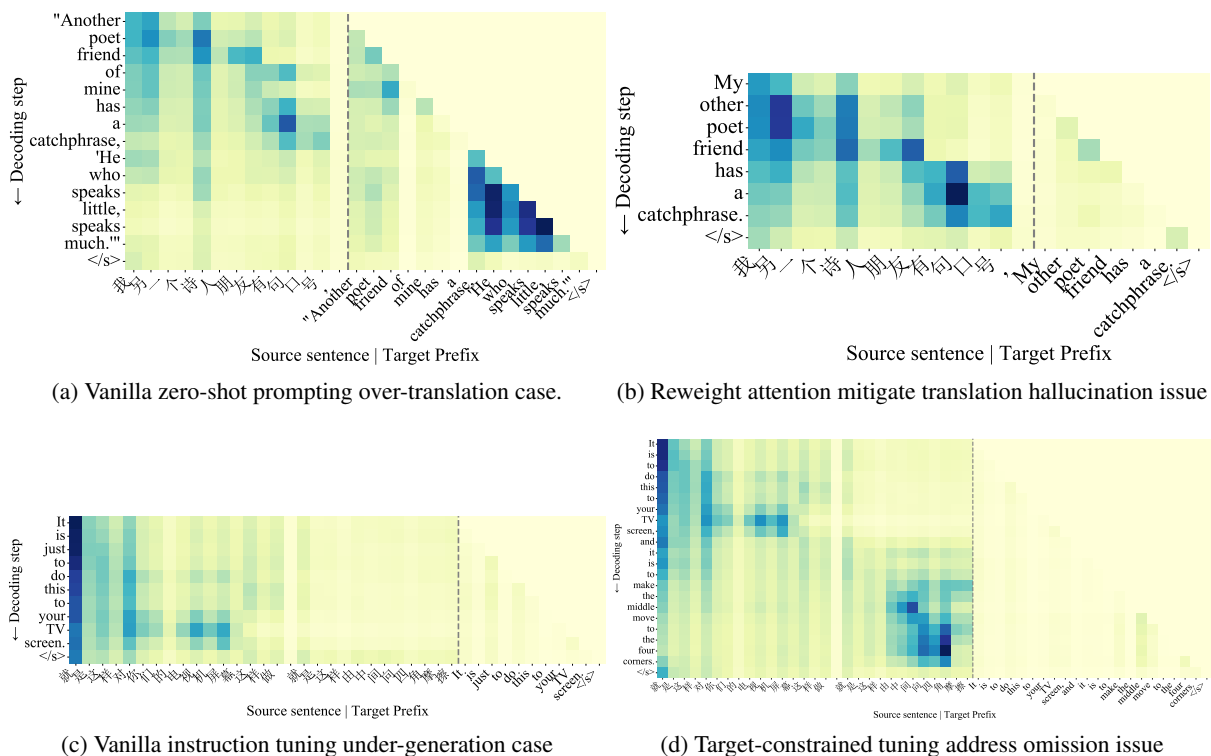(d) Target-constrained tuning address omission issue

Figure 5: Contribution visualization for unfaithful translations and corresponding mitigation across various settings. One of the correct translations of the first row is: "*Another poet friend of mine has a slogan.*"; One of the correct translations of the second row is: "*Just do this to your TV screen, rub it from the center to the four corners.*"

4.2 Case Study Using ALTI+

For qualitative analysis, we present several hallucinated examples under different settings and corresponding mitigation instances from Chinese to English (Zh ⇒ En) translation direction in Figure 5. As expected, hallucination spans ought to be discernible in our contribution visualization method, either as an emphasis on specific target prefixes or as a drop in the contribution of the source sentence, as manifested in Figure 5a and Figure 5c. Compared to the vanilla zeroshot baseline, the reweight attention approach can accurately direct the model's attention to the source context and generate source-aligned translation, thereby mitigating the hallucination issue as depicted in Figure 5b. It's crucial to emphasize that our reweight attention strategy doesn't impact the segments that are already accurately translated. It enhances the translation process when LLMs find it challenging to confidently pay attention to the source context by facilitating concentration on the relevant source tokens while generating the translation. When it comes to vanilla instruction tuning, the phenomenon of omissions frequently occurs in its generation(e.g., in Figure 5c), indicating that LLMs still struggle to cover the entire source context in translations, even after supervised tuning. By implementing target-constrained tuning, we ensure that LLMs do not excessively depend on the target prefix, but rather exploit as much of the source context as possible. This approach fosters the generation of more accurate and source-related translations than vanilla instruction tuning, as demonstrated in Figure 5d.

5 Related Works

5.1 Improving coverage of neural sequence to sequence model

Over the past few years, the coverage for attention mechanism and hallucinations in neural machine translation(NMT) have been the subject of study for several years (Tu et al., 2016; Shan et al., 2021; Li et al., 2022; Lee et al., 2019; Müller et al., 2020; Dale et al., 2023). Tu et al. (2016) suggested the inclusion of coverage information to provide additional data about the probability of source words. Mi et al. (2016) introduced explicit coverage embedding models to mitigate issues of unfaithful translation in NMT. Tu et al. (2017) proposed a dynamic mechanism to regulate the ratios at which source and target contexts contribute to

the generation of target words via context gates. Fu et al. (2023) identified the attention degeneration problem, i.e., as the generation step number grows, less and less attention is focused on the source sequence, in language models. Numerous methods have been proposed to enhance faithfulness in natural language generation (Li et al., 2022), all of which were designed to address unfaithfulness under the encoder-decoder architecture neural network. With the emergence of decoder-only LLMs, our focus shifted to the coverage of LLM-based translation. Although these approaches were originally designed for encoder-decoder models and cannot be directly applied to LLMs, they provide a valuable guide for current work on decoder-only language models (Chen et al., 2023; He et al., 2024). The research most similar to ours is that of Chen et al. (2023). They also focus on the unfaithful translation of LLMs, but their methods of addressing these problems differ significantly from ours. They enhance instruction comprehension by adding an instruction representation to the subsequent input and response representations, which are tailored to the tuning application scenario. In contrast, our methods fit different application scenarios. Reweighting attention works in low-resource settings without a full parallel corpus. Contrastive decoding fits API-only LLMs without internal information access. Target-instruction tuning suits tuning scenarios without adding extra modules or parameters.

5.2 LLMs for machine translation

Recently, several studies focused on how to prompt LLMs for machine translation (Zhang et al., 2023a; Vilar et al., 2023; Agrawal et al., 2023). Zhu et al. (2023) evaluated the performance of eight prominent large language models, including ChatGPT and GPT-4 (OpenAI et al., 2023), in multilingual machine translation. Jiao et al. (2023b) conducted a preliminary evaluation of ChatGPT's machine translation abilities, exploring translation prompts, multilingual translation, and robustness. ChatGPT competes well with commercial products in high-resource European languages but faces challenges with low-resource or distant languages. Parrot (Jiao et al., 2023a), a framework for enhancing chat-based translation, utilizes open-source LLMs, human-written translations, and feedback data to address challenges posed by restricted APIs. Xu et al. (2023) proposed a novel fine-tuning approach for LLMs that is specifically designed for the trans-

lation task and achieves significant advancement on LLM-based machine translation compared to the previous attempts. While LLMs have showcased remarkable performance, they inevitably confront various challenges in practical applications, with hallucinations emerging as one of the most notable issues (Wang et al., 2023; Zheng et al., 2023; Zhang et al., 2023b). Guerreiro et al. (2023) examine hallucinations in large multilingual translation models, conducting a comprehensive analysis on both conventional M2M neural machine translation models (Fan et al., 2021) and ChatGPT, shedding light on the unfaithfulness translation brought by LLMs.

However, previous studies have not explored the relationship between unfaithful or hallucinatory translation and the contribution from input tokens to generated tokens in LLMs. We have attempted to fill this gap. Our analysis of the contribution scores of input tokens has allowed us to highlight the limitations inherent in the decoder-only architecture of LLMs, specifically the lack of an explicit cross-attention module. Such a restriction can lead to inadequate focus on the source context, which we believe might potentially increase the risk of unfaithfulness in LLMs' generations.

6 Conclusion

In this research, we identify the issue of insufficient focus on source context in LLMs when applied to machine translation tasks and accordingly, we propose the reweight attention to adjust the attention weight of source context to help models focus on the source context during generation, contrastive decoding to reduce the influence of target prefixes, and target-constrained tuning to encourage LLMs to avoid excessive dependence on specific target prefixes. Our experimental results show marked improvements in translation performance across several language pairs in our proposed unfaithful translation test sets, outperforming baseline methods and effectively reducing the phenomenon of hallucinatory and unfaithful translations. Both our quantitative and qualitative analysis of contribution scores indicate the significance of our proposed methods in addressing the identified issue. While we only explore related issues in the application of machine translation, it is natural to extend our methods to other seq2seq tasks(e.g., summarization), which we leave for future exploration.

Acknowledgment

Limitations

This research conducts a preliminary investigation into the hallucinatory and unfaithful translations resulting from insufficient focus on the source context in decoder-only LLMs, and the following aspects can be improved upon in future work:

- **Variations in Instructions:** In our study, we did not consider the effects of varying the instruction prompt, nor did we examine the impact of our proposed methods under different instructions.

- **Testing Limited to Greedy Search:** Despite the availability of numerous decoding strategies and generation configurations, we set the temperature to 0 and 'do_sample' to False to demonstrate our proposed methods' effectiveness. We did not investigate the performance of these methods in combination with other generation strategies, such as beam search, top-k sampling, or nucleus sampling.

- **Increased Latency** Although our proposed methods are effective, they incur a higher computational cost compared to the standard settings. The contribution-based alignment selection strategy requires extra time to compute the dot product. Target-constrained tuning necessitates two forward passes, nearly doubling the training time.

Ethical Considerations

In our human-collected datasets, translations are obtained using zero-shot prompting from open-sourced LLMs, and thus any problematic responses can be attributed to the organizations that release these LLMs. We do not anticipate any significant risks associated with our research. In theory, our framework for mitigating unfaithful or hallucinatory translation could yield higher-quality translations without any toxic content. Based on our observations, our proposed methods have not resulted in any detrimental responses. To ensure the reproducibility of our experiments, we intend to make our code and evaluation data available to the public.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

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*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. 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.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107):1–48.

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.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. Transactions of the Association for Computational Linguistics, 11:1500–1517.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. Transactions of the Association for Computational Linguistics, 12:229–246.

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.

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 International Conference on Learning Representations.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. ParroT: Translating during chat using large language models tuned with human translation and feedback. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 15009–15020, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? yes with gpt-4 as the engine.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In Proceedings of the 2021 Conference on Empirical Methods in Natural

Language Processing, pages 4547–4568, Online and Punta Cana, Dominican Republic. 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.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods.

xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960, Austin, Texas. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak

Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

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.

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

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 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.

Yong Shan, Yang Feng, and Chenze Shao. 2021. Modeling coverage for non-autoregressive neural machine translation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. *arXiv e-prints*, page arXiv:2305.14739.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan

13828

Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

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, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

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.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Analyzing Contribution of input tokens

In each layer, the attention block computations can be expressed simply as a linear function of the input representations. Given a model with H heads, the attention block output of the i-th token y_i is computed by applying the layer normalization (LN) over the sum of the residual vector x_i and the output of the multi-head attention module (MHA) $\hat{x}_i$.

$$y_i = \text{LN}(\hat{x}_i + x_i) \qquad (9)$$

After the MHA module, $\hat{x}_i$ can be expressed as the linear combination of different input tokens and different attention heads:

$$\hat{x}_i = \sum_h^H W_o^h \sum_j^J A_{i,j}^h W_v^h x_j + b_o \qquad (10)$$

Given a vector u, $\text{LN}(u)$ can be reformulated as $\frac{1}{\sigma(u)} L u + \beta$, where L is a linear transformation. By swapping summations and utilizing the linearity of LN, we can now rewrite Eq. 9 as:

$$y_i = \sum_j T_i(x_j) + \frac{1}{\sigma(\hat{x}_i + x_i)} L b_o + \beta \qquad (11)$$

where the transformed vectors $T_i(x_j)$ are:

$$\text{T}_i(x_j) = \begin{cases} \frac{1}{\sigma(x_i + x_j)} L\left(\sum_h W_o^h A_{ij}^h W_v^h x_j\right) & \text{if } i \neq j \\ \frac{1}{\sigma(\hat{x}_i + x_i)} L\left(\sum_h W_o^h A_{ii}^h W_v^h x_j + x_i\right) & \text{if } i = j \end{cases} \qquad (12)$$

Kobayashi et al. (2020, 2021) propose assessing the contribution of each input vector x_j to the layer output y_i through the Euclidean norm: $c_{i,j} = \|T_i(x_j)\|_2$. While Ferrando et al. (2022a) argue that transformed representations exhibit reduced anisotropy and they suggest using l_1 norm: $c_{i,j} = \|y_i - T_i(x_j)\|_1$. Normalizing these contributions yields a layer-wise contribution matrix $C \in R^{J \times J}$. By employing a similar method as attention roll-out (Abnar and Zuidema, 2020), an overall contribution matrix for input tokens to the generated token y_i is obtained.

While previous studies have investigated the contributions of input tokens in machine translation (Ferrando et al., 2022b), their focus was exclusively on encoder-decoder style transformers. To the best of our knowledge, we are the first to analyze the contribution of input tokens in LLMs. Taking into account both the orientation and norm of the transformed vectors, we modify $c_{i,j} = \frac{T_i(x_j) \cdot y_i}{\|y_i\|_2}$ to represent the contribution of

transformed vectors towards the generated tokens. A larger vector projection onto y_i is expected to indicate a higher contribution. Given that future tokens are masked in the Transformer decoder, there is an inherent bias toward the initial tokens of the input sequence. Direct aggregation of each layer-wise contribution matrix may further intensify this bias (Abnar and Zuidema, 2020). Therefore, to prevent this, we normalize each layer-wise contribution matrix before aggregation.

B Experimental Setting

B.1 Instruction prompts

Translation prompt for LLaMA2-chat model. Our translation approach builds upon the work of Sennrich et al. (2024) The input to Llama2-chat consists of a system prompt and an instruction. To ensure that the assistant's response begins with the actual translation rather than an introductory phrase or prologue, we force-decode the prefix of the assistant response. Here is the zeroshot translation instruction prompt:

<s>[INST] «SYS»

You are a machine translation system that translates sentences from English to German. You just respond with the translation, without any additional comments.

«/SYS»

Sie stehen keine 100 Meter voneinander entfernt: Am Dienstag ist in Gutach die neue B 33-Fußgängerampel am Dorfparkplatz in Betrieb genommen worden - in Sichtweite der älteren Rathausampel.
Translate to English [/INST]Sure, here's the translation:

In the case of fewshot translation prompt, we adopt the same fewshot strategy used in Zhu et al. (2023). We use eight randomly sampled translation pairs from the respective training set as in-context exemplars. These exemplars are presented in "<X>=<Y>" format, where "<X>" and "<Y>" are the placeholder for the source and target sentence. Line-break serves as the exemplar's concatenation symbol.

Similar to traditional translation systems, we use bilingual sentence pairs to instill basic translation capabilities into LLMs. We adopt the Stanford Alpaca method (Taori et al., 2023) to convert bilingual sentence pairs into an instruction-following format, which fine-tunes LLMs for translation tasks. For the instruction tuning prompt

of the LLaMA2-chat model, we retained the zeroshot translation instruction prompt, but we did not compute the loss for the instruction query part. Instead, we only calculated the loss for the response output label, as illustrated in the subsequent examples of the translation instruction prompt:

<s>[INST] «SYS»
You are a machine translation system that translates sentences from English to German. You just respond with the translation, without any additional comments.
«/SYS»
Zwei Anlagen so nah beieinander: Absicht oder Schildbürgerstreich?
Translate to English [/INST]Sure, here's the translation:Two sets of lights so close to one another: intentional or just a silly error?

B.2 Evaluation Data

Unfaithful Translation Data. Specifically, we use our contribution scores analysis tool adapted for LLMs, which is modified from the ALTI+[5] method, to filter the data. If the source text contributions minus the target prefixes' contributions fall below a certain threshold, we collect them. We apply our methods to filter evaluation data on publicly available parallel data, such as News-Commentary v16 for German to English (De⇔En) and TED2013 for Chinese to English (Zh⇔En)[6]. After applying these criteria, we obtain 1009, 1002, 1010, and 1010 evaluation data sets for the De⇒En, En⇒De, Zh⇒En, and En⇒Zh tasks, respectively. Given our choice of translation instances with low source token contributions, which results in a dataset that contains instances that either deviate from the original sentence or lack semantic connection (e.g., copied instructions, text continuation, hallucinatory translation).

General Data. We evaluate the translation performance of LLMs on two sources of test sets:

- **Flores-101:** We use the Flores-101 which serves as the evaluation benchmark for multilingual-machine systems and the number of test samples is 1012 for all translation directions (Goyal et al., 2022).

- **WMT22 Test sets:** We also utilize the test sets from the WMT22 competition. These sets are constructed based on recent content from various domains, including news, social, e-commerce, and conversational domains. The sample numbers for the De⇒En, En⇒De, Zh⇒En, and En⇒Zh tasks are 1984, 2037, 1875, and 2037, respectively (Kocmi et al., 2022).

B.3 Model Training

We conduct our main experiments with Hugging-Face Transformers[7] on open-source LLMs from the LLaMA2 family (Touvron et al., 2023). Specifically, we choose LLaMA2-7b-chat[8], BLOOMZ-7b1-mt[9] (Muennighoff et al., 2023), ChatGLM3-6b[10] (Du et al., 2022), and Vicuna-7b[11] (Chiang et al., 2023) with matched parameters, and also include LLaMA2-7b-chat, LLaMA2-13b-chat and LLaMA2-70b-chat to study the effect of model sizes. The hyperparameters used for finetuning are mainly aligned with those of Stanford Alpaca[12] (Taori et al., 2023). For instruction tuning and target-constrained tuning, we finetune models over 5 epochs with a learning rate of 1e-4, using the corresponding language direction parallel data. During the implementation of the LoRA finetuning[13], we set the 'lora_r' to 16, 'lora_alpha' to 32, 'lora_dropout' to 0.3. The 'target_modules' are configured to include the query, key, value projection and output projection within the attention module. For more specific hyperparameters, please refer to our released scripts. We perform the finetuning process on 8 Nvidia A100 GPUs and employ DeepSpeed[14] ZeRO stage 2 for model parallel.

C Ablation Study

We analyze specific factors related to our proposed methods that may impact the translation performance of LLMs. As a default setting, we perform ablation studies on our proposed unfaithful translation dataset using the LLaMA2-7b-chat model.
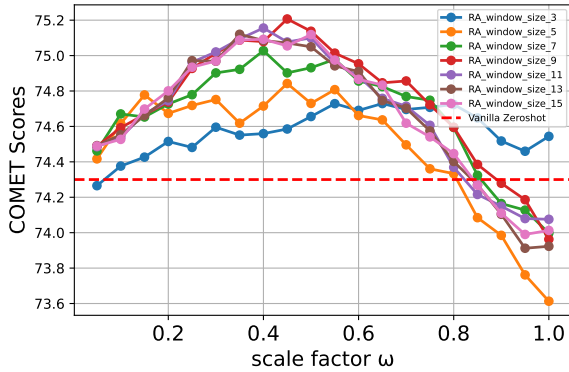
Figure 6: COMET scores over scale factors ω and window size D parameters of reweight attention

Figure 7: BLEU scores over the various penalty degrees α

C.1 Reweight Attention different strategy and the effect of the scale factor ω and window size

strategy	BLEU	COMET
baseline zeroshot	11.4	74.3
Monotonic Local Window	11.8	74.7
Heuristic Local Window	12.5	75.0
Global Window	12.4	75.0

Table 5: different strategies for the reweight attention

Two strategies for choosing local window anchors to adjust attention scores. For comparison, we have also introduced a global window strategy that adjusts the attention scores across the entire source context. In this study, we aim to determine which strategy most effectively improves the translation performance of LLMs. As shown in Table 5, our heuristic strategy, which prioritizes the most significant source tokens based on their contribution, outperforms the other strategies. The monotonic strategy is the most subtle optimization strategy, as aligning the source and target tokens in a step-by-step manner may not conform to a human-like translation. Interestingly, adjusting the attention scores across the entire window of source tokens produces results comparable to the heuristic approach. We further investigated the translation performance across varying local window sizes and scale factor ω, as shown in Figure 6. Our results suggest that significant performance enhancement is observed with larger windows when the window size is initially small(e.g. 3 or 5). However, as the window size increases beyond a certain point(e.g. 9), we noti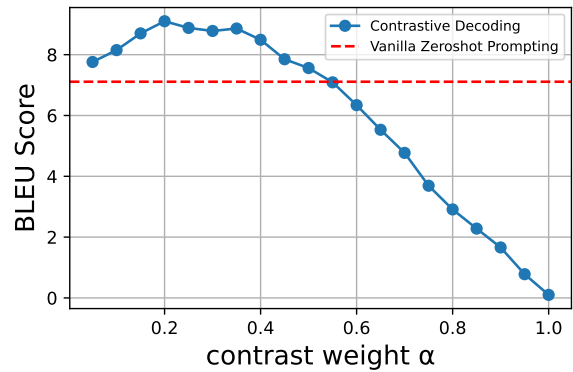ce a slight decrease in translation performance. This can be attributed to the fact that incorporating irrelevant source information, not related to the currently generated token, may introduce additional noise. Additionally, assigning a small factor does not effectively shift the model's focus toward the source tokens. Conversely, overemphasis on the source context can lead to performance decline. In total, the optimal balance appears to be a window size of around 9 and a scale factor of around 0.5.

C.2 Effect of adjustment level α in Contrastive Decoding

In our proposed contrastive decoding method, we introduce a hyperparameter α to control the penalty degree of contrastive decoding. A smaller α results in the model predictions' output distribution being closer to the original distribution of the next tokens. We carry out experiments using various values of α and present the results in Figure 7. Smaller values of α(e.g., 0.1) do not yield performance as robust as larger values α(e.g., 0.5), suggesting that the models continue to generate unconditioned tokens when a penalty of a lower degree is utilized. However, increasing the value of α(e.g., 1) further causes a decline, due to the unfluent and ungrammatical generation from erasing too much target contribution.

C.3 Effect of mask ratio β and KL-divergence coefficient λ in target-constrained tuning

We study the influence of two parameters: the mask ratio β and the weight assigned to the KL-divergence loss λ. Here, we let β vary between $\{0 \to 0.5\}$ and λ between $\{0 \to 1\}$ and From the results in Figure 8 and 9, we find that: 1) A mask ratio of 0.15 and a KL coefficient of 0.5 yields the best performance; 2) Target-constrained
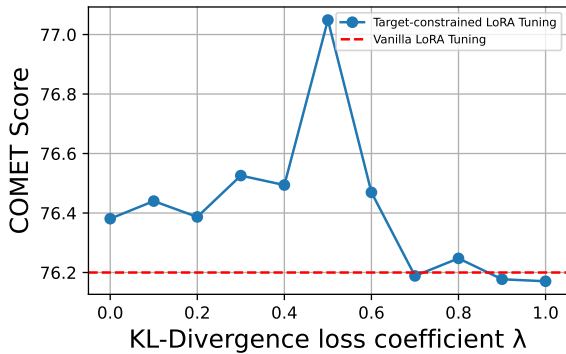
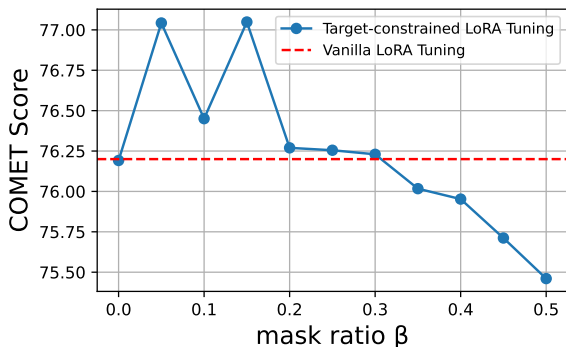Figure 8: COMET scores over different the KL Divergence coefficients λ

Figure 9: COMET scores over different mask ratios β

tuning consistently achieves strong results with a mask ratio 0.05 and 0.15; 3) As the mask ratio increases, target-constrained tuning fails to converge because the model struggles with the high randomness from the masked target prefix; 4) Smaller λ values underperform compared to larger ones, highlighting the need for attention to KL-divergence regularization. However, over-regularization is detrimental. The best balance is achieved at $\lambda = 0.5$.

D Human Evaluation

In this appendix, we describe the manual evaluation. First, we detail the simple guidelines that were presented to manual annotators. Second, we report the number of annotators and inter-annotation agreements.

Fleiss's Kappa Scores		
Correct	Hallucination	Errors
0.95	0.79	0.88

Table 6: Fleiss's Kappa inter-annotator agreement scores for the three annotation categories.

annotation guidelines The annotators were provided with the guidelines as outlined in Table 7. We consolidate the labels based on a majority vote. Notably, we also designate unfaithful translations, encompassing under-translations and over-translations, as hallucinations. Translations that exhibit semantic deviation from the source context are also recognized as hallucinations. Concretely, for the purpose of reporting, we grouped 'Named-entity mistranslation' and 'Off-target' under the 'Error' category, while 'Semantically detached', 'Omission', and 'Over-translation' were classified under the 'Hallucination' category."

Inter-annotation agreement To ensure the reliability and quality of our annotated unfaithful translation datasets, we additionally assigned the same set of another 100 randomly sampled translation instances to two annotators. The Fleiss' Kappa statistic, representing the agreement in the assessment of annotation categories between the annotators, is presented in Table 6. As demonstrated, we can effectively classify the different types of translation with strong agreement between the two annotators, thereby indicating the effectiveness of the human evaluation test set. This confirms the suitability of our annotated data for our analysis.

E Connection between Human Evaluation and Automatic metrics

To reveal the consistency of hallucination evaluation between automated metrics and human evaluation in cases of unfaithful translation, we conduct further analysis. This analysis focuses on the correlation between human evaluations quantified by numerical scores and automated metrics. According to Table 7 in Appendix D, translations are first categorized into one of six annotation categories, and the detailed categories of the sampled set are shown in Table 8. To convert these categories into numerical scores, we can use the proportion of hallucinatory translations to measure hallucination severity, following Dale et al. (2023). A higher proportion indicates more severe hallucinations. Alternatively, to emphasize unfaithfulness in translation, we assign a score of -0.4 for omission and over-translation, a score of -0.25 for semantically detached, a score of -0.2 for error translations, and a score of +1 for correct translations.

We first convert human evaluation categories into numerical scores using the previously dis-

Annotation Types	Definition
Correct	The translation fully conveys the meaning of the original text. It may contain content that does not affect the availability of the content or Minor understandability errors (eg: incorrect punctuation conversion)
Omission	Translation is an example of omission if and only if part of the source language sentence is translated correctly but the remaining part is not (not including an attempt to translate that part but the translation is wrong, which is a complete lack of attempt) to translate that part of the content)
Semantically Detached	Some of these incorrect translations are supported by the content of the source language sentences. However, a large proportion of this mistranslation is not (it conveys a different meaning than the one in the source sentence).
Off-target	An example of off-target translation is when the translation system fails to translate the source language into the target language, that is, non-target language fragments appear in the translation.
Named-Entity	Named-entities mistranslation are mistranslated (for example: mistranslation of names of people, places, organizations, dates, prices, etc.)
Over-translation	A translation is an example of over-translation if and only if all the contents of the source language sentence are translated correctly, but the translation system excessively generates more translations.

Table 7: Human annotations Guidelines

Method	semantically detached	omission	over-translation	off-target	name-entity
Vanilla zeroshot	28	18	18	9	6
Reweight attention	11	8	6	8	6
Contrastive decoding	12	11	3	8	6
Vanilla instruction tuning	15	3	3	3	6
Target-constrained tuning	8	1	0	3	5

Table 8: The detailed categories of the collected 100 examples in Section3.

Method	BLEU	COMET	BLEURT	Human Evaluation Numerical Score
Vanilla zeroshot	7.45	65.5	49.3	-14.2
Reweight attention	11.3	68.7	53.1	43.65
Contrastive decoding	11.7	69.2	52.6	40
Vanilla instruction tuning	14.8	74.7	58.9	57.85
Target-constrained tuning	15.4	75.0	60.5	70.5

Table 9: The numerical scores of human evaluation and automatic metrics on the sampled set

cussed method, and then compute BLEU, COMET, and BLEURT (Sellam et al., 2020) metrics for these examples. The results are presented in the Table 9

Finally, we calculated the Pearson correlation coefficient between the automated metrics and human evaluation scores. The Table 10 shows that the BLEU metric aligns most closely with human evaluation scores, followed by COMET. Thus, we further confirm that BLEU and COMET are proper for our experiments as they can reflect the level of unfaithful translation to a certain degree.

Automatic_metric	Pearson Correlation Coefficient
BLEU	0.9640
COMET	0.8942
BLEURT	0.8907

Table 10: The Pearson correlation coefficient between automatic metrics and human evaluation numerical score.

Contribution Metric	Unfaithful dataset	WMT22	Flores101
Llama2-7b-chat average contribution score	0.7914	0.8381	0.8434

Table 11: The contribution metric scores between different test sets.

mean	min	max	median	mode	std
39.9	7	300	33	22	26.94

Table 12: The basic statistics of the source length in the Zh-to-En test set.

F Analysis of the imbalanced improvement phenomenon

Our paper focuses on unfaithful translations in LLMs caused by inadequate attention to the source context. Therefore, we collect data containing such issues to form a specific dataset and conduct main experiments on this dataset, with results in Table 1 showing our methods effectively address the unfaithful translation issue. Only then did we generalize our methods to open general datasets, with results in Table 2. Experimental results show improvement of our method is less significant on general datasets than on the unfaithful translation test set. This is mainly due to the rare unfaithful translations when using the Llama-2-7b-chat model for zero-shot translation on the Flores101 and WMT22 test sets. We also calculate source contribution scores, as detailed in Appendix A (higher scores mean more focus on the source context during LLM generation, thus is less possible to generate unfaithful translations), across both general and proposed test sets to illustrate this point. As shown in the Table 11, the source contribution scores on the general dataset are higher than the proposed dataset, which explains why our method's improvements are less significant on general datasets.

G Further analysis of the sudden drop

The lines in the figure represent averages not for all examples but for those corresponding to examples with at least i tokens. Therefore, when the number of tokens exceeds 100, only translations with a sentence length greater than 100 tokens are taken into account. The basic statistics of the source length in the Zh-to-En test set are presented in Table 12.

We conducted the human evaluation to the long source context cases, and found that when the source context length exceeds 200, translations

produced by LLM using vanilla zero-shot settings show more instances of not following instructions and more omissions, resulting in reduced reliance on the source context during translation generation. Consequently, the contribution of the source context diminishes.

H Extended comparison experiments between model families and scales within WMT22 test set and unfaithful test set

In the comparison of these language model families, ChatGLM3-6b demonstrates superior performance in the English to Chinese (En $\Rightarrow$ Zh) language translation direction, largely as a result of its enhanced modeling of Chinese during the pretraining phase. On the other hand, BLOOMZ-7b1-mt performs better in the Chinese to English (Zh $\Rightarrow$ En) language translation direction. This enhancement can be ascribed to its substantial exposure to a varied compilation of parallel multilingual pretraining corpora, coupled with the prevalence of English in the pretraining corpus of BLOOM.

As evidenced by the results shown in Table 16, the 7b model experiences the most significant boost, while larger models only achieve a marginal enhancement. Given that the test unfaithful translation sets are collected from the poorest translation instances of the 7b chat model, larger models demonstrate superior translation capabilities and generate fewer unfaithful issues within such a dataset. This inconsistent improvement is primarily due to the difference in the number of unfaithful issues generated by LLMs of varying scales.

System	En ⇒ De		De ⇒ En		En ⇒ Zh		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base Model:BLOOMZ-7b1-mt								
Vanilla Instruction tuning LoRA	20.5	64.2	31.8	76.6	44.6	84.4	25.5	79.1
Target-constrained tuning LoRA	21.2	64.4	32.5	76.7	45.2	84.6	26.2	79.4
Base Model:ChatGLM3-6b								
Vanilla Instruction tuning LoRA	27.1	73.2	37.6	81.0	46.4	84.4	26.8	79.1
Target-constrained tuning LoRA	27.2	73.2	37.8	81.1	46.7	84.5	27.2	79.2
Base Model:Vicuna-7b								
Vanilla Instruction tuning LoRA	32.5	80.6	39.8	82.0	42.0	82.6	25.0	76.8
Target-constrained tuning LoRA	33.5	81.3	41.4	83.1	42.9	83.0	25.2	76.9

Table 13: Translation performance of various families of LLMs with a similar size on WMT22 test sets.

System	En ⇒ De		De ⇒ En		En ⇒ Zh		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base Model:LLaMA2-7b-chat								
Vanilla Instruction tuning LoRA	34.0	81.7	42.0	83.2	38.9	81.6	25.5	78.3
Target-constrained tuning LoRA	34.5	81.8	42.5	83.4	39.2	81.8	25.8	78.4
Base Model:LLaMA2-13b-chat								
Vanilla Instruction tuning LoRA	37.0	83.2	43.0	83.5	43.7	84.2	27.6	79.3
Target-constrained tuning LoRA	37.6	83.4	43.8	83.9	43.9	84.2	27.8	79.4
Base Model:LLaMA2-70b-chat								
Vanilla Instruction tuning LoRA	41.3	84.8	46.0	84.7	49.2	85.6	30.2	80.5
Target-constrained tuning LoRA	41.8	85.0	46.9	84.9	49.9	86.0	31.1	80.8

Table 14: Translation performance of various families of LLMs with a similar size on WMT22 test sets.

System	En ⇒ De		De ⇒ En		En ⇒ Zh		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base Model:BLOOMZ-7b1-mt								
Vanilla Instruction tuning LoRA	11.9	62.7	21.1	75.3	20.9	78.3	16.2	77.5
Target-constrained tuning LoRA	12.0	62.7	21.9	76.0	21.2	78.4	17.1	77.8
Base Model:ChatGLM3-6b								
Vanilla Instruction tuning LoRA	15.4	70.2	25.8	80.4	20.5	78.5	16.3	77.2
Target-constrained tuning LoRA	15.9	70.3	26.2	80.5	21.3	78.9	17.0	77.5
Base Model:Vicuna-7b								
Vanilla Instruction tuning LoRA	15.6	69.9	24.6	78.7	20.5	77.7	14.2	71.6
Target-constrained tuning LoRA	16.0	70.1	25.8	79.5	21.1	77.9	14.8	71.7

Table 15: Translation performance of various families of LLMs with a similar size on human-collected unfaithful translation test sets.

System	En ⇒ De		De ⇒ En		En ⇒ Zh		Zh ⇒ En	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base Model:LLaMA2-7b-chat								
Vanilla Instruction tuning LoRA	18.6	77.4	28.7	79.8	18.6	76.2	15.5	76.5
Target-constrained tuning LoRA	20.0	77.9	30.1	81.1	19.1	76.5	16.6	77.0
Base Model:LLaMA2-13b-chat								
Vanilla Instruction tuning LoRA	21.3	80.4	30.7	81.5	20.0	77.8	17.7	77.9
Target-constrained tuning LoRA	21.9	80.6	31.3	81.7	20.6	78.0	18.2	78.1
Base Model:LLaMA2-70b-chat								
Vanilla Instruction tuning LoRA	23.1	81.4	33.4	84.5	21.7	79.1	19.7	79.0
Target-constrained tuning LoRA	23.8	81.8	33.9	84.8	22.1	79.3	20.4	79.3

Table 16: Translation performance of various families of LLMs with a similar size on human-collected unfaithful translation test sets.