Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models

Anonymous ACL submission

Abstract

In Neural Machine Translation (NMT), models will sometimes generate repetitive or fluent output that is not grounded in the source sentence. This phenomenon is known as hallucination and is a problem even in large-scale multilingual translation models. We propose to use Contrastive Decoding, an algorithm developed to improve generation from unconditional language models, to mitigate hallucinations in NMT. Specifically, we maximise the log-likelihood difference between a model and the same model with reduced contribution from the encoder outputs. Additionally, we propose an alternative implementation of Contrastive Decoding that dynamically weights the difference based on the maximum probability in the output distribution to reduce the effect of CD when the model is confident of its prediction. We evaluate our methods using the Small (418M) and Medium (1.2B) M2M models across 21 low and medium-resource language pairs. Our results show a 14.6 ± 0.5 and 11.0 ± 0.6 maximal increase in the mean COMET scores for the Small and Medium models on those sentences for which the M2M models initially generate a hallucination., respectively.

1 Introduction

Hallucinations are a rare but problematic phenomenon in NMT (Neural Machine Translation) whereby the target side output is repetitive or fluent but not grounded in the source sentence (Ji et al., 2023). Even though hallucinations are rare in NMT, they are a significant problem as they undermine trust in deployed NMT systems. Hallucinations occur when the target side sentence is detached from the source side sentence (Wang and Sennrich, 2020; Raunak et al., 2021; Dale et al., 2023), or in other words, when there is a low contribution of the source sentence to the generation of the target sentence.

Previous work on mitigating hallucinations has focused on sampling translations and reranking them according to quality metrics (Dale et al., 2023; Guerreiro et al., 2023b). Separate to this, Li et al. (2022) proposed Contrastive Decoding (CD) as a way of mitigating bad behaviour (such as excessive repetition and low diversity) when generating from unconditional language models. CD is a decoding algorithm that maximises the difference between the log probabilities of a strong expert and a weak amateur model (equivalent to maximising the ratio of probabilities). A threshold is applied so that decoding follows the expert when it is more confident. The intuition behind CD is that the amateur model is more prone to certain types of low-quality generation, so by subtracting the log probabilities, these are removed. We hypothesise that by using CD with an amateur, which is prone to source detachment, we can mitigate hallucinations in NMT.

In order to create an amateur with low source attachment, we experiment with different strategies for reducing the role of cross-attention. The simplest is the NO ENCODER strategy, where the amateur is a decoder-only version of the expert. In our other strategies, we retain the encoder and cross-attention but impose uniform attention, remove attention from the most highly attended source position, or scale down all cross-attention values.

In contrast to unconditional generation, NMT should be more strictly grounded in the source sentence. Additionally, hallucinations only account for a small proportion of translations, and hence, mitigation of hallucinations must not come at the cost of reduced performance on other sentences. As such, increasing the diversity of the translations is less desirable than it is in unconditional generation. Ideally, CD would only take effect when a model is hallucinating. To address this issue, we experiment with a novel variant of CD that dynam-

ically adjusts the subtraction's magnitude based on a distribution's maximum value.

We evaluate our approach on large multilingual models, which have recently been shown to be prone to hallucinations (Guerreiro et al., 2023a). Specifically, we use the M2M family of models (Fan et al., 2020) and consider the 418M (Small) and 1.2B (Medium) versions.

We summarise our contributions as follows:

- We show that using CD in conjunction with amateur models that have reduced source contributions mitigates hallucinations.
- We extend the CD algorithm, dynamically setting the weight given to the amateur to limit the effect of CD when the expert is confident.
- We evaluate across 21 language pairs using the M2M family of models on the FLORES-101 dataset, reporting a mean increase of 14.6 ± 0.5 and 11.0 ± 0.6 COMET on sentences causing hallucinations for the Small model and Medium models respectively.

2 Related Work

Hallucination Detection We discuss Hallucination detection as it relates to our experimental setup. Guerreiro et al. (2023b); Dale et al. (2023); Raunak et al. (2021) all evaluate different methodologies for automatically identifying hallucinations and demonstrate the effectiveness of ALTI+ as a hallucination detection method.

Hallucination Mitigation Guerreiro et al. (2023a) propose using a different fallback model when a hallucination is detected. Other methods rely on sampling and reranking translations, for example, using COMET (Rei et al., 2022) to mitigate hallucinations (Guerreiro et al., 2023b; Dale et al., 2023). Compared to our methodology, this approach relies on an additional outside model to rank sentences to achieve the best performance. Additionally, both works only evaluate on a small de→en model, whereas we evaluate on large-scale multilingual models.

In contemporaneous work, Sennrich et al. (2023) uses a similar CD approach to mitigate hallucinations. Hallucinations are evaluated by counting the proportion of segments with chrF2 <10 (Lee et al., 2019; Müller and Sennrich, 2021). Unlike our work, the authors use the same model as the amateur but supply it with randomly selected inputs. In contrast, we use different models as the amateur, supplied with the same inputs. Randomly selecting another source segment is potentially less stable than using a model as an amateur, as translations can depend heavily on the selection of the source segment. This work also compares different amateurs and techniques for combining the expert and amateur distributions, whereas Sennrich et al. (2023) places additional focus on off-target translations. Our work can thus be seen as complementary to theirs.

3 Methodology

We first describe CD as proposed by Li et al. (2022), then discuss our proposed improvements (normalisation and dynamic weighting), and finally motivate the amateur models that we use in our experiments.

3.1 Contrastive Decoding

Equation 1 gives the ORIGINAL formulation of CD in log space proposed by Li et al. (2022). Here $p_x(i)$ is the probability (post softmax) assigned to token i in vocabulary $\mathcal{V}$ by the expert model, and $p_a(i)$ is the probability assigned by the amateur model.

$$CD(i) = \log(p_x(i)) - \gamma \log(p_a(i)) \quad (1)$$

The subtraction results in a new set of scores $CD(i)$ that are used in beam search in place of the expert's scores. γ is a hyperparameter that weights the amateur subtraction. The equivalent formulation in linear space equates to rescaling the expert probabilities according to the amateur probabilities.

Li et al. (2022) use a hyperparameter α ($0 < \alpha < 1$) to threshold the expert probability distribution. As shown in Equation 2, only those tokens with a probability greater than or equal to the maximum probability scaled by hyperparameter α are considered for CD.

$$\mathcal{V}_{thresh} = \{i \in \mathcal{V} : \log(p_x(i)) \geq \log(\alpha) + \max_j \log(p_x(j)\} \quad (2)$$

As stated by (Li et al., 2022) this thresholding has two purposes. Firstly, preventing extremely unlikely tokens under the expert being the highest scoring under CD and secondly, if the expert is significantly confident, to consider only one token so that CD selects the same token as the expert.

3.2 Contrastive Decoding for Hallucinations

Unlike Li et al. (2022) our aim is not to increase the diversity of generations but rather to prevent hallucinations in NMT. This presents two fundamental challenges. Firstly, hallucinations only represent a small proportion of translated sentences, and secondly, compared to open-ended generation, the output of NMT needs to be grounded in the source sentence. As such CD when applied in our context, should ideally only affect the output for those few sentences where the model hallucinates and only minimally change those outputs that are not hallucinations. We address these issues through normalisation and dynamic weighting.

3.3 Normalisation

The motivation of normalisation is both to stabilise beam search when decoding and also to help with the dynamic weighting only approach introduced subsequently. Without normalisation, the magnitudes of the CD scores at each time step are different, and as a result, time steps will contribute differently to the hypotheses in the beam. As dynamic only applies CD to a significant number of tokens at each time step, normalisation helps in the case where a certain beam shows a much bigger CD score than other beams.

$$N_{CD} = \frac{\sum_{i=1}^{\mathcal{V}_{thresh}} p_x(i)/p_a(i)^{\gamma}}{\sum_{i=1}^{\mathcal{V}_{thresh}} p_x(i)} \quad (3)$$

Equation 3 gives the value of the normalisation constant, used to normalise the CD probabilities. Dividing the contrastive scores by N_{CD} normalises the scores before scaling them to sum to the probability mass covered by $\mathcal{V}_{thresh}$. The set of normalised CD scores is combined with the set of expert probabilities given by the complement of $\mathcal{V}_{thresh}$ to obtain a probability distribution. We refer to CD with normalisation as NORMALISED

3.4 Dynamic Weight

The original CD algorithm is applied at each time step with the same weight, and hence, all probabilities are rescaled. Rather than varying the number of candidates CD considers (as the threshold α does), we propose to vary the degree to which CD affects token generation by dynamically setting the γ in Equation 1. Equation 4 gives the dynamic weighting approach to setting the amateur weight γ, where β is a hyperparameter.

$$\gamma = 1 - \max_i p_x(i)^{\beta} \quad (4)$$

When the expert distribution has a high maximum probability $\gamma \rightarrow 0$, thereby reducing the effect of CD when the expert model is confident. Conversely, when the expert distribution has a low maximum probability $\gamma \rightarrow 1$ and, hence, the rescaling due to the amateur probabilities is larger. We experiment with both a combination of thresholding and dynamic weighting (DYNAMIC); and solely relying on dynamic weighting by setting the threshold in Equation 2 so that the number of tokens considered for CD is constant (DYNAMIC ONLY).

3.5 Amateur Models

In order to mitigate hallucinations, amateur models are chosen to simulate detachment from the source, thereby stimulating hallucinations or at least increasing the probability mass assigned to hallucinated tokens while decreasing the probability mass of "reasonable" tokens. Apart from the SMALL amateur, the different approaches all try to reduce the source contribution to the output:

NO ENCODER: The No Encoder approach calls the decoder without the encoder inputs (bypassing the entire cross-attention block), essentially acting as a language model. Without the source sentence, the amateur has to rely only on the target side prefix. Unlike Language Model fusion (Stahlberg et al., 2018), which interpolates the two distributions, CD rescales the distributions, increasing scores that are unlikely under the amateur and decreasing scores that are likely under the amateur.

FLAT ATTENTION: An amateur where the cross-attention scores are uniform, which equates to taking the unweighted mean of the encoder outputs. In this approach, the amateur still has access to the encoder information with only the attention information removed, representing a softer detachment.

ZERO MAX ATTENTION: For this approach, we set the maximum cross-attention score of the amateur to zero, and hence, there is no contribution from the most salient encoder output. When selecting the maximum, we disregard the last token to account for punctuation at the end of a sentence.

ATTENTION SCALING: This approach is used both independently and in combination with the Flat Attention and Zero Max Attention approaches. We directly reduce the contribution of the source by scaling down all of the attention weights.

SMALL: Using a smaller model trained on the same data (Li et al., 2022). We use the M2M Small amateur as a comparison against the other amateurs that explicitly reduce the source contribution.

3.6 Beam Search

After calculating the CD scores, we use beam search to generate the translations. As all beams are the same for the first time step, we ensure that when not using normalisation, we set the threshold probability to the probability of the tenth token, ensuring that all beams have a valid score. As the normalised scores include the output probabilities of the expert, this is no longer necessary for normalised CD.

4 Experimental Setup

4.1 Models and Datasets

We adopt the setup of Guerreiro et al. (2023a) and use the M2M (Fan et al., 2020) models, evaluating on the FLORES-101 (Goyal et al., 2022) dataset. The M2M models are strong multilingual transformer (Vaswani et al., 2017) models that are trained on 7.5B sentences but still have been shown to hallucinate for low and medium resource languages (Guerreiro et al., 2023a), thus providing a tested method to evaluate our approach. We only evaluate on the Small (418M) and Medium (1.2B) M2M models, as they produce more hallucinations than the 12B parameter model. The language pairs we evaluate on are given in Table 1 alongside the number of detected hallucinations on the expert. For evaluation, we combine the FLORES-101 dev and devtest splits to increase the number of hallucinations. All our experiments are run using fairseq (Ott et al., 2019)[1].

4.2 Metrics

Hallucination Detection As both WMT and FLORES-101 do not have gold standard labels for hallucinations, we follow Guerreiro et al. (2023a) and use a combination of ALTI+ (Ferrando et al., 2022) and TNG (top n-gram count) (Raunak et al., 2021; Guerreiro et al., 2023b) to detect hallucinations in the expert translations. ALTI+ measures both the source and the target contribution to generations and can be used to identify detached sentences. We use the same approach as Guerreiro et al. (2023a) and obtain a threshold value for the ALTI+ score using en→de and en→ru WMT-19 (Barrault et al., 2019) and en→fr WMT-14 data (Bojar et al., 2014) data. TNG identifies sentences where the top target side n-gram count is at most t greater than the top source side n-gram, where n is set to 4 and t is set to 2. Additionally, reasonable quality thresholds for spBLEU (Goyal et al., 2022), chrf++ (Popović, 2015), and COMET (Rei et al., 2022) are used to filter out false positives. We report all threshold values in Appendix A.

Evaluation Metrics: We report COMET (Rei et al., 2022) scores as these have been shown to be sensitive to hallucinations (Guerreiro et al., 2023b; Dale et al., 2023) for our main results[2]. We evaluate how CD affects hallucinations by splitting our test sets into hallucinations and non-hallucinations and report COMET for each separately. Additionally, we report hallucination counts for our selected approaches using the hallucination detection pipeline detailed above. As ALTI+ is a model-based metric, we use the expert with forced decoding to generate ALTI+ scores for the translations generated with CD.

4.3 Hyperparameters

We decode using Beam search with a Beam size of 4. We tune the following hyperparameters: α, γ and β. We determine hyperparameters for the M2M experiments by performing a grid search using ha→en WMT-21 data (Akhbardeh et al., 2021)[3] as it has a reasonable number of hallucinations. Hyperparameters were selected using the maximum COMET score on EXPERT hallucinations. For the DYNAMIC ONLY approach, we fix the number of tokens used for CD to 25 for all experiments; the Attention Scaling parameter is set to 0.01 when used independently and 0.25 when combined with other approaches. The complete set of hyperparameters is given in Appendix A.

5 Results

First, we present results comparing our different experimental setups by combining all language pairings. We compare amateur models using the ORIGINAL CD approach before reporting on the effects of our additions to the CD algorithm. Next, we compare the performance across languages by looking at the distributions of COMET scores and presenting qualitative examples. Finally, we use the hallucination detection suite to report the num-

[1] https://github.com/facebookresearch/fairseq

[2] Specifically, we use `wmt22-comet-da`

[3] WMT data obtained using SACREBELU (Post, 2018)

Language Pair	ast-en	en-ast	oc-en	en-oc	ps-en	en-ps	sw-en	en-sw	bn-en	en-bn	fa-en	en-fa	tr-en	en-tr	zh-en	en-zh	be-ru	fr-sw	ar-fr	el-tr	hi-bn
Small (418M)	6	111	5	30	35	1465	2	195	6	169	3	2	1	3	1	5	190	514	3	5	117
Medium (1.2B)	33	29	2	92	11	1109	2	54	0	62	1	16	0	3	0	2	99	431	6	7	93

Table 1: Language pairs used for evaluation along with hallucination counts detected for the EXPERT models.

	Small(418M)				Medium(1.2B)			
	ORIGINAL	NORMALISED	DYNAMIC	DYNAMIC ONLY	ORIGINAL	NORMALISED	DYNAMIC	DYNAMIC ONLY
EXPERT	47.4±0.4	-	-	-	54.4±0.5	-	-	-
ATTENTION SCALING	**62.0**±0.3	59.5±0.3	60.8±0.4	60.7±0.4	**65.4**±0.4	61.6±0.5	62.2±0.5	61.3±0.5
FLAT ATTENTION	55.5±0.4	56.4±0.4	54.3±0.4	54.9±0.4	57.4±0.5	58.6±0.5	56.5±0.5	57.2±0.5
FLAT ATTENTION SCALING	61.9±0.3	60.1±0.4	60.8±0.4	61.3±0.4	64.2±0.4	61.6±0.5	**62.3**±0.5	**62.8**±0.5
NO ENCODER	**62.0**±0.3	**61.1**±0.3	**61.7**±0.4	**61.6**±0.4	64.9±0.4	**62.3**±0.5	62.1±0.5	62.8±0.5
SMALL	-	-	-	-	54.2±0.4	58.1±0.5	57.6±0.5	57.8±0.5
ZERO MAX ATTENTION	59.3±0.4	58.1±0.4	57.4±0.4	58.0±0.4	60.5±0.5	59.3±0.5	58.6±0.5	59.8±0.5
ZERO MAX ATTENTION SCALING	60.9±0.3	60.0±0.4	60.1±0.4	60.8±0.4	63.5±0.4	62.0±0.5	61.7±0.5	61.9±0.5

Table 2: Mean COMET scores on examples with hallucination for the two M2M models used (Small, Medium). The mean is calculated over sentences in all translation directions (en→ps and ps→en are removed due to having far more hallucinations than all other language pairs). Errors reported are SEM (Standard Error on the Mean). Bold and underlined values highlight the maximum in each column and row.

ber of detected hallucinations when using CD. For completeness, we also report spBLEU, chrF++ and COMET for our selected approach in Appendix B.

5.1 Amateur Models and Dynamic Weighting

Contrastive Decoding Reduces Hallucinations Evaluating on M2M models presents a robust multilingual experimental setup that evaluates across 21 low and medium-resource language pairs. We compare our experimental approaches by splitting the test sets into hallucinations and non-hallucinations and averaging the COMET scores across all language pairs. Table 2 shows that all variants of CD increase the mean COMET scores for both the Small and Medium M2M models, confirming our hypothesis that CD - when using an amateur designed to hallucinate - generates improved translations of sentences for which the EXPERT generates hallucinations. The results for the Medium model show a maximal increase in the mean COMET-22 of 11.0 ± 0.5. In contrast, the Small model shows a maximal increase of 14.6 ± 0.6, suggesting that either CD is better for the smaller model or more likely that the hallucinations for the 1.2B parameter model are less severe.

Amateurs that remove the most encoder information are better at reducing hallucinations We first focus on evaluating the performance of the amateur models in terms of mean COMET scores for hallucinations with the ORIGINAL approach. NO ENCODER, ATTENTION SCALING, FLAT ATTENTION SCALING, and ZERO MAX ATTENTION SCALING all achieve comparable results, as shown in the first column of Table 2, when using the ORIGINAL CD approach. We hypothesise that the M2M models have a strong enough decoder that FLAT ATTENTION and ZERO MAX ATTENTION do not remove enough encoder information to promote

	Small(418M)				Medium(1.2B)			
	ORIGINAL	NORMALISED	DYNAMIC	DYNAMIC ONLY	ORIGINAL	NORMALISED	DYNAMIC	DYNAMIC ONLY
EXPERT	77.4±0.1	-	-	-	80.8±0.1	-	-	-
ATTENTION SCALING	74.3±0.1	73.1±0.1	75.7±0.1	**77.0**±0.1	**78.2**±0.1	77.4±0.1	79.6±0.1	**80.6**±0.1
FLAT ATTENTION	**74.9**±0.1	**75.3**±0.1	74.7±0.1	75.0±0.1	78.0±0.1	78.9±0.1	78.5±0.1	78.7±0.1
FLAT ATTENTION SCALING	74.2±0.1	75.1±0.1	75.7±0.1	75.8±0.1	78.0±0.1	**79.1**±0.1	79.6±0.1	79.8±0.1
NO ENCODER	74.2±0.1	74.8±0.1	**76.4**±0.1	75.8±0.1	**78.2**±0.1	78.7±0.1	**80.1**±0.1	79.8±0.1
SMALL	-	-	-	-	71.8±0.1	77.1±0.1	78.7±0.1	80.1±0.1
ZERO MAX ATTENTION	73.7±0.1	75.2±0.1	74.9±0.1	75.1±0.1	77.5±0.1	79.0±0.1	78.9±0.1	79.0±0.1
ZERO MAX ATTENTION SCALING	73.5±0.1	74.9±0.1	75.7±0.1	75.9±0.1	77.7±0.1	79.0±0.1	79.7±0.1	79.8±0.1

Table 3: Mean COMET scores of non-hallucinations for the two M2M models used (Small, Medium). The mean is calculated over sentences in all translation directions (en→ps and ps→en are removed due to having far more hallucinations than all other language pairs). Errors reported are SEM (Standard Error on the Mean). Bold and underlined values highlight the maximum in each column and row.

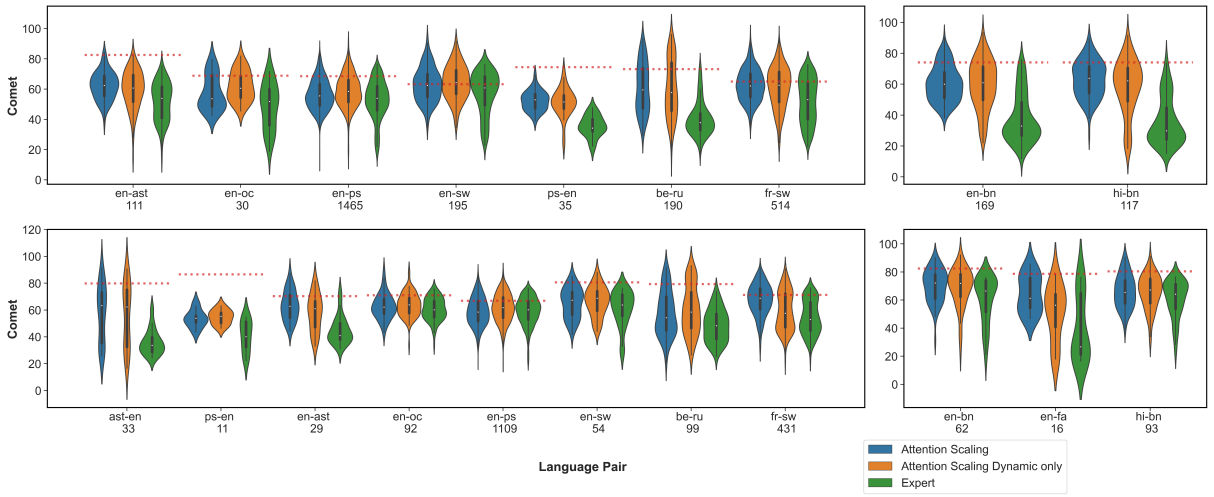

Figure 1: COMET score distributions for hallucinations using the EXPERT and ATTENTION SCALING amateurs for all language pairs with 10 or more hallucinations. Results are given for both the 418M and 1.2B parameter models. The red line is the mean COMET of EXPERT non-hallucinations for a given language pair.

hallucinations in the other amateurs. As FLAT ATTENTION SCALING and ZERO MAX ATTENTION SCALING both rely on being combined with ATTENTION SCALING we prefer the more straightforward approaches of removing the encoder outputs or simply scaling the attention weights as amateurs.

For the medium model, we also experiment with using the Small 418M parameter model as an amateur. Using the ORIGINAL approach, SMALL does not improve mean COMET scores for examples with hallucinations compared to the expert. Although smaller models tend to have lower source contribution, the fact that SMALL does not improve on the expert lends further credence to our claim that actively restricting the source information available to amateur models improves the ability of CD to correct expert hallucinations.

Based on the FLAT ATTENTION ZERO MAX ATTENTION and SMALL results, we propose that removing encoder information is important when mitigating hallucinations using CD.

Dynamic Weighting mitigates the adverse effect of CD on non-hallucinations Looking at Table 3, we see that the ORIGINAL implementation of CD adversely affects sentences for which the EXPERT does not hallucinate. The results demonstrate that our DYNAMIC weighting approaches counteract the adverse effects of CD on these sentences, achieving comparable COMET scores to the EXPERT model. We underline this result by reporting the mean chrF++ scores for non-hallucinations in Table 4 as COMET scores are robust against paraphrasing. As chrF++ is a string-based metric, the decreased results highlight that the ORIGINAL CD approach generates translations which are less similar to the reference than the EXPERT and the DYNAMIC or DYNAMIC ONLY approaches. As we define our task as mitigating hallucinations, we argue that these decreases are undesirable.

	Small(418M)		Medium(1.2B)	
	ORIGINAL	DYNAMIC ONLY	ORIGINAL	DYNAMIC ONLY
EXPERT	45.0±0.1	-	50.3±0.1	-
ATTENTION SCALING	42.5±0.1	**46.3**±0.1	**45.9**±0.1	**49.4**±0.1
FLAT ATTENTION	**43.4**±0.1	43.9±0.1	45.6±0.1	47.3±0.1
FLAT ATTENTION SCALING	42.2±0.1	44.6±0.1	45.7±0.1	48.6±0.1
NO ENCODER	42.4±0.1	44.6±0.1	45.8±0.1	48.6±0.1
SMALL	-	-	38.5±0.1	**49.4**±0.1
ZERO MAX ATTENTION	43.0±0.1	44.6±0.1	46.2±0.1	48.4±0.1
ZERO MAX ATTENTION SCALING	41.9±0.1	45.0±0.1	45.6±0.1	49.0±0.1

Table 4: Mean chrF++ scores for non-hallucinations on the two M2M models used (Small, Medium). The mean is calculated over sentences in all translation directions (en→ps and ps→en are removed due to having far more hallucinations than all other language pairs). Errors reported are SEM (Standard Error on the Mean).

The effects of DYNAMIC weighting are particularly pronounced for SMALL for which the chrF++ score in Table 4 is 11.8 ± 0.1 less than the EXPERT without DYNAMIC weighting. Looking at Table 2, the COMET of SMALL DYNAMIC increases, indicating that the SMALL with DYNAMIC weight corrects some EXPERT hallucinations. Taken together, these results show that scaling the weight by the maximum probability prevents CD from making significant changes to EXPERT translations whilst still affecting translations to mitigate hallucinations.

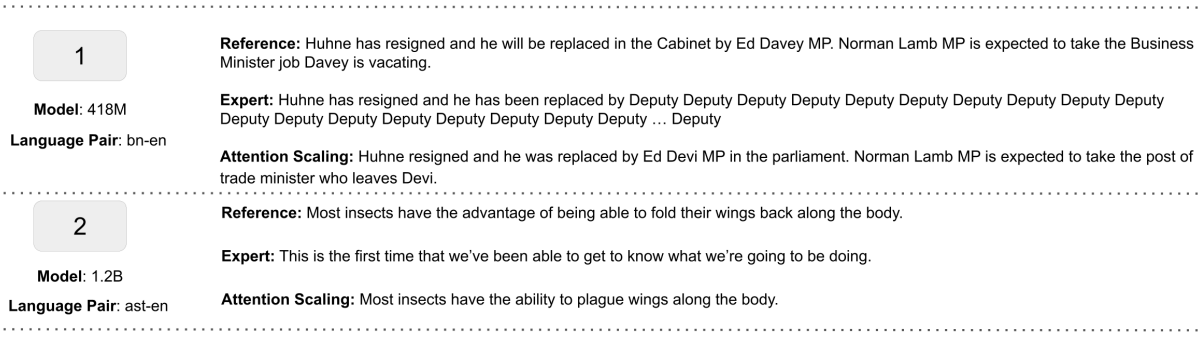

Figure 2: Examples EXPERT hallucinations that are fixed by ORIGINAL CD with the ATTENTION SCALING amateur.

5.2 Attention Scaling

Next, we explore how CD behaves across different languages and present examples to showcase how CD repairs EXPERT hallucinations. For this we select the ATTENTION SCALING ORIGINAL and ATTENTION SCALING DYNAMIC approaches. Whilst, NO ENCODER has comparable COMET scores, and if anything appears more robust, we select ATTENTION SCALING as it is more flexible by allowing fine-grained selection of the source contribution via a hyperparameter. However, we believe both amateurs would have been viable choices. We report on DYNAMIC ONLY to better understand how its behaviour contrasts with the ORIGINAL approach.

Repaired hallucinations still have errors In Figure 2 we provide examples of sentences for which the expert generates hallucinations which CD fixes. Compared to the EXPERT both sentences generated with CD are not detached from the source. The fixed detachment is especially apparent in sentence 2, which contains the incorrect segment 'ability to plague' but recovers and correctly translates 'along the body'. Both sentences still have errors, but we argue that these are translation errors rather than hallucinations, in line with the topology proposed by Guerreiro et al. (2023b). Another way of looking at these errors is that CD highlights tokens that rely on the source, as these should have a lower score in the amateur when the source context is removed. However, the expert still needs to assign enough probability to such tokens for them to be considered in the first place. Hence, in such cases, CD still promotes the correct tokens, but the model is not able to translate the source sentence correctly.

Repairing hallucinations is dependent on the quality of the expert model We analyse COMET scores across languages by plotting the distribution of COMET scores for EXPERT hallucinations in Figure 1, along with the mean COMET for non-hallucinations. The result shows that CD with ATTENTION SCALING increases the COMET of EXPERT hallucinations across all languages. Additionally, for all languages apart from ast-en with the Medium model, CD improves the long tail of comet scores observed for hallucinations. The mean COMET scores for non-hallucinations show that they act as a ceiling to the improvement of hallucinated translations. For some language pairs, such as hi-bn (Small), the entire COMET score distribution improves as the EXPERT model has the potential to generate better translations. In contrast, for weaker translation directions such as en-ps and en-sw, Figure 1 illustrates that most COMET scores do not improve because the overall performance of the model caps them; instead, only the long tail is improved. We propose that the ability of CD to fix

Language Pair	Experiment	ast-en	en-ast	oc-en	en-oc	ps-en	en-ps	sw-en	en-sw	bn-en	en-bn	fa-en	en-fa	tr-en	en-tr	zh-en	en-zh	be-ru	fr-sw	ar-fr	el-tr	hi-bn
Small (418M)	EXPERT	6	111	5	30	35	1465	2	195	6	169	3	2	1	3	1	5	190	514	3	5	117
	ATTENTION SCALING DYNAMIC ONLY	0	57	1	15	6	1238	1	90	5	47	1	0	0	1	1	5	93	358	3	0	40
	ATTENTION SCALING	0	20	0	3	0	1111	0	41	2	4	0	1	0	0	0	1	70	171	0	0	5
Medium (1.2B)	EXPERT	33	29	2	92	11	1109	2	54	0	62	1	16	0	3	0	2	99	431	6	7	93
	ATTENTION SCALING DYNAMIC ONLY	15	10	0	58	1	800	1	39	0	46	0	6	0	2	0	2	65	360	2	3	73
	ATTENTION SCALING	15	0	0	34	2	508	0	13	0	15	0	0	0	2	0	2	57	162	1	1	27

Table 5: Hallucination counts for the EXPERT model and the ATTENTION SCALING ORIGINAL and ATTENTION SCALING DYNAMIC ONLY approaches. Hallucination labels are obtained with the ALTI+ and TNG hallucination detection methodology.

hallucinations depends on the strength of the EXPERT model for a given language pair. We support this claim by observing that, in Table 5, both en-ps and en-sw have a high proportion of hallucinations that CD does not repair.

CD with Attention Scaling fixes translations In Table 5 we present the hallucination counts using the ATTENTION SCALING and ATTENTION SCALING DYNAMIC ONLY obtained using the hallucination detection pipeline. As ALTI+ is a model-based metric, we force decode the CD translations using the expert to obtain ALTI+ scores, hypothesising that for expert hallucinations that are repaired, the ALTI+ scores improve based on the target prefix. The table shows that CD with ATTENTION SCALING reduces hallucinations for all language pairs. Adding DYNAMIC ONLY decoding reduces the efficacy of CD, with fewer hallucinations being mitigated. However, we observe in Figure 1 that the DYNAMIC ONLY approach has higher peak COMET scores for the small model. This leads us to speculate that when DYNAMIC weighting fixes a hallucination, it may generate better translation.

Model	Small (418M)		Medium (1.2B)	
Hallucination	Yes	No	Yes	No
Amateur Temperature				
0.5	58.5±0.4	71.9±0.1	61.3±0.5	76.0±0.1
1	**62.0**±0.3	74.3±0.1	**65.4**±0.4	78.2±0.1
1.5	61.4±0.5	**75.2**±0.1	63.0±0.6	**79.1**±0.1

Table 6: Mean COMET scores across all language pairs for the ATTENTION SCALING amateur using different Softmax temperatures for the amateur model.

The performance of CD depends on the amateur temperature As Li et al. (2022) demonstrate that the amateur temperature can affect CD performance, we report the mean COMET scores for hallucinations and non-hallucinations at different amateur temperatures in Table 6. We can see that reducing the temperature to 0.5 and thereby increasing the sharpness of the amateur distribution degrades the COMET scores for non-hallucinations and hallucinations. By contrast, increasing the amateur temperature to 1.5 decreases the mean COMET for hallucinations but slightly increases the COMET for non-hallucinations. This result makes sense as increasing the temperature leads to a smoother distribution.

6 Conclusions

This paper applies CD decoding to the task of mitigating hallucinations in NMT. We show that CD improves the mean COMET scores of sentences for which the M2M translation models generate hallucinations. Our results also support our hypothesis that a key part of effectively using CD to mitigate hallucinations is restricting decoder access to the encoder outputs in amateur models, simulating target detachment from the source.

Additionally, we experiment with decreasing the adverse effect of CD on sentences for which the M2M models already generate good translations by dynamically changing the weight hyperparameter, which scales the subtraction of the amateur probabilities. We show that dynamic weighting decreases the changes to translations generated compared to the expert, but this comes at the cost of repairing fewer hallucinations. Improvements to the dynamic approach would require a model-based metric that identifies hallucinations at the token level.

As such, we recommend the original approach and either removing the encoder outputs or scaling down the cross-attention weights. In light of the adverse effects on non-hallucinations, we also suggest using CD only when a hallucination is detected, for example, with a hallucination detection pipeline. Any 'fixed' hallucinations should also be flagged if CD is used in a deployed system. End users should be made aware that the translation was originally a hallucination and may still contain translation errors.

7 Limitations

Whilst we evaluate across 21 language pairs, these are all medium and low-resource languages. We provided no results on how our method works with high-resource languages. Our experimental setup does not investigate out-of-domain translations where hallucinations are particularly frequent. We also point out that we fix the number of tokens considered by the DYNAMIC ONLY approach rather than trying different values. Finally, our implementation for NO ENCODER skips the entire cross-attention block. As such, the associated layer normalisation is also skipped. Hence, the results of NO ENCODER and ATTENTION SCALING with the hyperparameter set to 0 do not lead to the same translations, but we do not investigate this further.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

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, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. ArXiv:2010.11125 [cs].

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. 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.

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 Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in Large Multilingual Translation Models. ArXiv:2303.16104 [cs].

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. 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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

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. 2022. Contrastive Decoding: Open-ended Text Generation as Optimization. ArXiv:2210.15097 [cs].

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding 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 259–272, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2023. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium. Association for Computational Linguistics.

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

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

A Hyperparameters

We first use an en→de model that has a data set of human-annotated hallucinations (Guerreiro et al., 2023b; Dale et al., 2023) in order to validate our approach, and explore different variants of our strategy.

Threshold	Value			
ALTI+ (Small)	ALTI+ (Medium)	spBLEU	chrF++	COMET
0.32	0.38	18.7	45.6	76.6

Table 7: Threshold values for the Hallucination Detection pipeline.

Experiment	Alpha(α)	Weight(γ/β)	Minimum Tokens	Scaling
ATTENTION SCALING	0.15	0.5	1	0.01
ATTENTION SCALING NORMALISED	0.05	1.0	1	0.01
ATTENTION SCALING DYNAMIC	0.01	0.5	1	0.01
ATTENTION SCALING DYNAMIC ONLY	-	0.2	25	0.01
FLAT ATTENTION	0.2	0.5	1	-
FLAT ATTENTION NORMALISED	0.01	0.5	1	-
FLAT ATTENTION DYNAMIC	0.01	0.7	1	-
FLAT ATTENTION DYNAMIC ONLY	-	0.7	0.25	-
FLAT ATTENTION SCALING	0.1	0.5	1	0.25
FLAT ATTENTION SCALING NORMALISED	0.01	0.5	1	0.25
FLAT ATTENTION SCALING DYNAMIC	0.01	0.5	1	0.25
FLAT ATTENTION SCALING DYNAMIC ONLY	-	0.5	25	0.25
NO ENCODER	0.15	1.0	1	-
NO ENCODER NORMALISED	0.1	1.0	1	-
NO ENCODER DYNAMIC	0.1	0.5	1	-
NO ENCODER DYNAMIC ONLY	-	0.2	25	-
SMALL	0.25	0.75	1	-
SMALL NORMALISED	0.1	0.75	1	-
SMALL DYNAMIC	0.05	0.5	1	-
SMALL DYNAMIC ONLY	-	0.2	25	-
ZERO MAX ATTENTION	0.1	0.5	1	-
ZERO MAX ATTENTION NORMALISED	0.01	0.5	1	-
ZERO MAX ATTENTION DYNAMIC	0.01	0.7	1	-
ZERO MAX ATTENTION DYNAMIC ONLY	-	0.7	25	-
ZERO MAX ATTENTION SCALING	0.1	0.5	1	0.25
ZERO MAX ATTENTION SCALING NORMALISED	0.01	0.5	1	0.25
ZERO MAX ATTENTION SCALING DYNAMIC	0.01	0.5	1	0.25
ZERO MAX ATTENTION SCALING DYNAMIC ONLY	-	0.5	25	0.25

Table 8: Hyperparameters used for all experiments. Alpha and weight were set using a grid search. We combine weight and dynamic weight parameters in one column. Minimum tokens refers to the minimum number of tokens for which CD is applied. Scaling refers to the magnitude of the ATTENTION SCALING used. Minimum tokens and scaling were set to the given values for all experiments.

Parameter	Values
Alpha(α)	0.01, 0.05, 0.1, 0.15, 0.2, 0.25
Weight(γ)	0.25, 0.5, 0.75, 1.0
Weight Dynamic(β)	0.2, 0.5, 0.7

Table 9: All hyperparameters that were tried as part of the grid search.

B Additional Results

	EXPERT			ATTENTION SCALING			ATTENTION SCALING DYNAMIC ONLY		
Language Pair	spBLEU	chrF++	COMET	spBLEU	chrF++	COMET	spBLEU	chrF++	COMET
ast-en	30.3	54.2	73.9	26.2	51.0	71.8	30.8	54.5	74.1
en-ast	24.8	48.5	68.0	18.1	45.1	66.6	24.2	48.6	68.8
oc-en	37.7	61.2	73.3	32.4	56.5	70.6	38.6	61.3	72.8
en-oc	23.3	47.5	68.2	15.6	43.1	66.6	22.0	47.2	69.2
ps-en	12.6	38.0	64.5	9.8	35.0	62.1	12.7	37.5	64.1
en-ps	5.8	22.6	55.7	3.6	21.0	57.7	5.1	23.0	60.3
sw-en	27.7	50.5	72.8	22.8	46.9	71.1	27.3	50.3	73.1
en-sw	21.4	46.3	72.9	16.3	42.9	70.3	20.8	46.2	73.3
be-ru	15.6	39.1	78.6	14.7	38.7	76.5	16.6	40.4	80.2
fr-sw	15.6	39.7	67.9	12.4	38.6	68.1	15.6	40.4	69.8
Mean Low-Resource	21.5	44.8	69.6	17.2	41.9	68.1	21.4	44.9	70.6
bn-en	25.6	51.3	82.7	19.8	47.2	80.9	24.9	50.9	82.6
en-bn	16.5	33.0	71.1	8.7	25.6	66.0	12.7	28.8	68.9
fa-en	28.2	53.9	82.3	22.0	49.4	80.4	27.7	53.6	82.5
en-fa	27.5	45.6	81.3	22.1	41.9	77.6	27.1	45.5	81.1
tr-en	31.4	55.7	84.7	24.9	51.0	82.9	30.9	55.4	84.8
en-tr	28.9	50.5	83.9	22.4	46.3	80.0	28.7	50.8	83.8
zh-en	21.7	48.4	81.8	17.0	44.5	80.1	21.7	48.1	81.9
en-zh	19.1	20.7	78.5	13.4	17.7	75.3	18.5	20.3	78.2
ar-fr	27.8	50.7	76.5	20.1	46.0	73.3	27.2	50.5	76.7
el-tr	19.8	42.5	79.5	15.0	39.4	76.2	19.8	42.8	79.7
hi-bn	16.2	32.7	71.9	8.8	25.3	68.2	13.6	29.8	71.3
Mean Medium-resource	23.9	44.1	79.5	17.7	39.5	76.4	23.0	43.3	79.2

Table 10: spBLEU, chrF++, COMET across all language pairs for the EXPERT, ATTENTION SCALING, and ATTENTION SCALING DYNAMIC ONLYexperiments with the Small(418M) model.

	Expert			Attention Scaling			Attention Scaling Dynamic Only		
Language Pair	spBLEU	chrF++	COMET	spBLEU	chrF++	COMET	spBLEU	chrF++	COMET
ast-en	36.8	58.7	79.2	30.7	54.6	77.5	36.5	58.6	79.4
en-ast	33.0	54.0	70.1	26.7	50.4	68.7	32.9	54.2	70.5
oc-en	46.4	66.9	79.1	39.7	62.0	77.1	46.9	66.7	79.0
en-oc	30.1	52.3	70.6	20.6	47.0	68.3	28.8	52.0	71.1
ps-en	17.9	43.4	71.2	14.0	39.8	68.4	17.4	42.5	70.5
en-ps	9.3	27.4	62.6	5.6	23.9	61.0	8.3	26.4	64.0
sw-en	35.2	57.0	79.7	29.4	52.9	78.1	34.6	56.6	79.7
en-sw	30.3	53.5	80.2	24.0	49.1	76.3	29.6	53.1	79.5
be-ru	19.3	42.5	84.8	17.7	41.3	82.9	19.4	42.8	85.2
fr-sw	21.5	44.5	74.0	18.5	43.6	74.3	21.3	44.6	74.4
Mean Low-resource	28.0	50.0	75.1	22.7	46.5	73.3	27.6	49.8	75.3
bn-en	28.4	53.1	83.9	22.2	48.9	82.4	28.0	53.0	83.9
en-bn	25.5	40.8	81.8	18.8	36.1	78.4	24.6	40.1	81.3
fa-en	30.0	54.5	82.8	23.3	49.8	81.0	29.2	54.0	82.8
en-fa	22.7	42.4	78.4	16.7	38.2	74.6	21.5	41.4	77.9
tr-en	35.7	58.7	86.7	28.7	53.8	85.4	34.8	58.1	86.7
en-tr	30.6	52.1	86.2	22.3	46.7	82.3	30.2	52.0	86.0
zh-en	27.1	52.2	84.6	21.0	47.7	83.0	26.4	51.7	84.6
en-zh	23.1	22.4	82.9	17.1	19.8	80.3	22.8	22.6	82.7
ar-fr	28.4	50.4	76.2	21.6	46.1	73.4	27.7	50.1	76.2
el-tr	21.7	44.1	82.0	15.7	40.1	79.0	21.5	44.3	82.2
hi-bn	23.6	38.4	79.5	18.1	35.0	77.3	23.4	38.5	79.6
Mean Medium-resource	27.0	46.3	82.3	20.5	42.0	79.7	26.4	46.0	82.2

Table 11: spBLEU, chrF++, COMET across all language pairs for the Expert, Attention Scaling, and Attention Scaling Dynamic Onlyexperiments with the Medium(1.2B) model.

C GPU Hours

All experiments where run on GTX 3090 GPUs. While we did not keep track of the GPU utilization we note that we only ran decoding experiments in this work. As an estimate of GPU hours both the parameter grid search and evaluation took 3 days running on 4 GPUs.