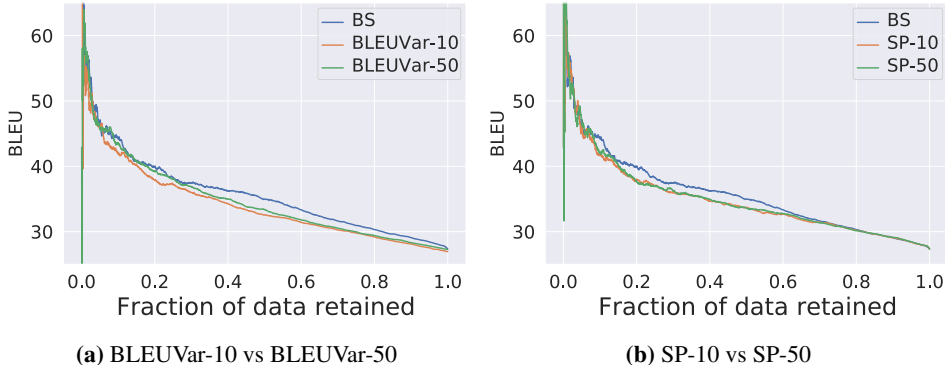


## A IN-DISTRIBUTION EXPERIMENTS

This section details our experiments on data that lays within the training distribution for the WMT English (EN)  $\rightarrow$  German (DE) and English (EN)  $\rightarrow$  Vietnamese (VI) tasks. We explore the calibration of Transformer models in this setting and evaluate the effectiveness of using MC Dropout and the proposed methods to measure the model uncertainty.

In addition to WMT13 dataset for EN  $\rightarrow$  DE tasks mentioned in the previous section, we use the **IWSLT 2015** dataset for translation tasks from EN to VI. There are 133k sentences pairs in the **IWSLT 2015** training set and 1.3k sentences pairs in the **IWSLT 2015** test set. Both the training and test data for **IWSLT 2015** come from the domain of TED talks.

### A.1 EVALUATING MODEL CALIBRATION



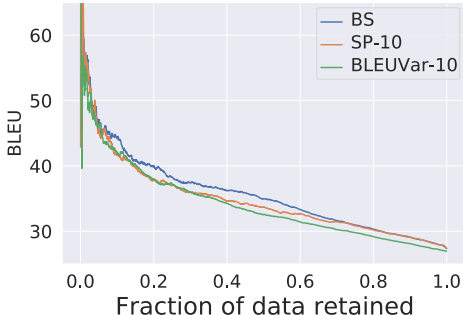
**Figure 4:** Uncertainty estimator comparisons for different number of samples. The model is trained for EN to DE tasks with 4.6m training data using 350k steps.

**Table 6:** AUC for plots in Figure 4 and Figure 5

BS	SP-10	SP-50	BLEUVar-10	BLEUVar-50
35.78	34.86	34.93	34.14	34.68

The first question we hope to answer is the quality of calibration in Transformers models and to evaluate the effectiveness of MC Dropout in improving uncertainty estimates.

The Transformer was trained on the full EN-DE training set (4.6 million samples) for 350k steps. We evaluate on the *newstest2014* test set.

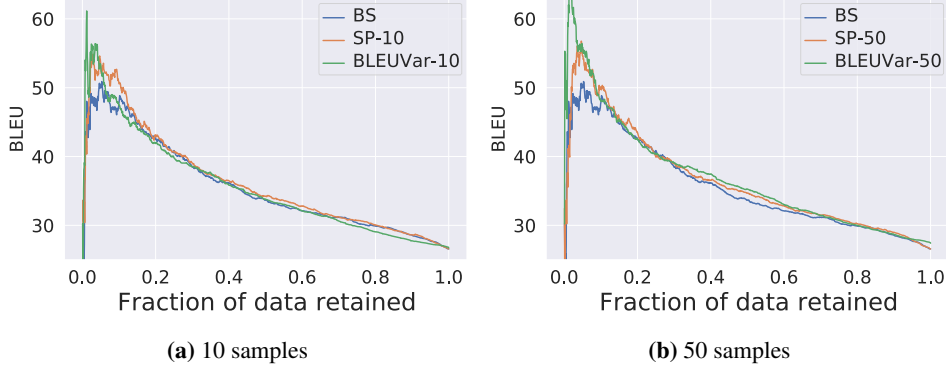


**Figure 5:** BLEU scores for different uncertainty estimators under various retained data rates. The model is trained for EN to DE tasks with 4.6m training data using 350k steps.

The results from Figures 4 and 5 suggest that the beam search score provides a well-calibrated uncertainty metric on the in-distribution test data. The second observation is that MC Dropout-based methods seem to slightly under-perform beam score in this setting (see Table 6), even when the

number of samples is increased fivefold. In this setting, our proposed metric (BLEUVar) benefits more from increasing the number of dropout samples relative to sequence probability.

## A.2 THE IMPACT OF TRAINING SET SIZE



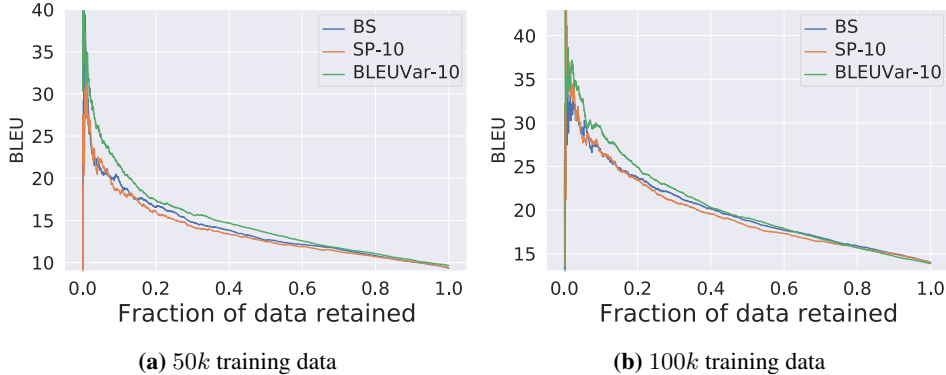
**Figure 6:** BLEU scores for different uncertainty estimators under various retained data rates. The model is trained for EN to VI tasks with 133k training data using 350k steps.

**Table 7:** AUC for plots in Figure 6

BS	SP-10	SP-50	BLEUVar-10	BLEUVar-50
35.55	36.25	36.33	35.66	<b>36.92</b>

The WMT EN-DE training set is fairly large and one would assume that most test sentences (or very similar ones) have been observed during training time. Hence we do not expect much epistemic uncertainty to exist in this testing scenario, which the experiments seem to confirm. A natural question to ask is on the effect of training set size on the calibration of models. We explore this question by considering the WMT English to Vietnamese (EN-VI) task which has 133k samples in the training set (approx. 2.6% of EN-DE), and down-sampling the EN-DE training set to 50k and 100k samples.

The performance-retention plots in Figure 6 and the AUC in Table 7 indicate that, while a large training set yields curves that seem to suggest beam score is a sufficient uncertainty metric, when a small dataset is used the MC Dropout-based uncertainty metrics begin to outperform the beam score (note the retention range 0.0 to 0.2). Moreover, in the small training set setting increasing the number of samples drawn from MC dropout results in a significant improvement for BLEUVar.

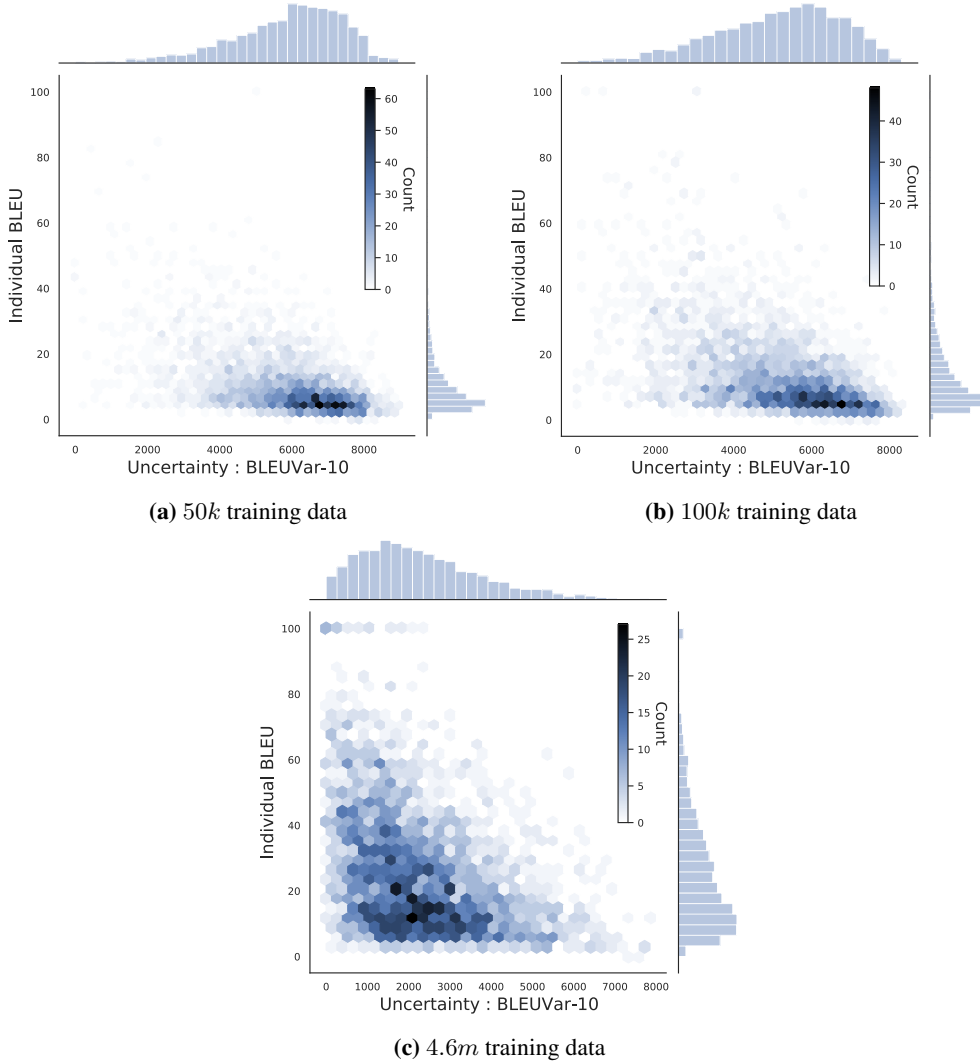


**Figure 7:** Uncertainty estimator comparisons for models with different sizes of training set. The models were trained for EN to DE tasks with 50k and 100k training data using 350k steps.

The experiments depicted in Figures 7 and 8 consist of down-sampling the EN-DE training set. Figure 7 and Table 8 demonstrates a similar pattern to the above EN-VI experiment when down-sampling the EN-DE data to 50k and 100k examples. Again, in the low-data regime BLEUVar substantially out-performs beam score and sequence probability. Figure 8 demonstrates the impact of data size

**Table 8:** AUC for plots in Figure 7

	BS	SP-10	BLEUVar-10
(a) 50k training data	13.97	13.61	<b>14.89</b>
(b) 100k training data	19.88	19.64	<b>20.44</b>



**Figure 8:** The density of individual BLEU score versus uncertainty (BLEUVar-10) for all sentences in the same test set *newstest2014* produced by models trained with various size of data set. The sentences are ordered by their uncertainty from low (left) to high (right) using BLEUVar-10. Following the calculation of BLEUVar, since we have 10 samples, the uncertainty estimate BLEUVar-10 has the value in range  $[0, 90]$ . And we scale it up by  $\times 100$ , which results in the x-axis has the range  $[0, 9000]$ . The models were trained for EN to DE tasks with 50k, 100k and 4.6m training data using 350k steps.

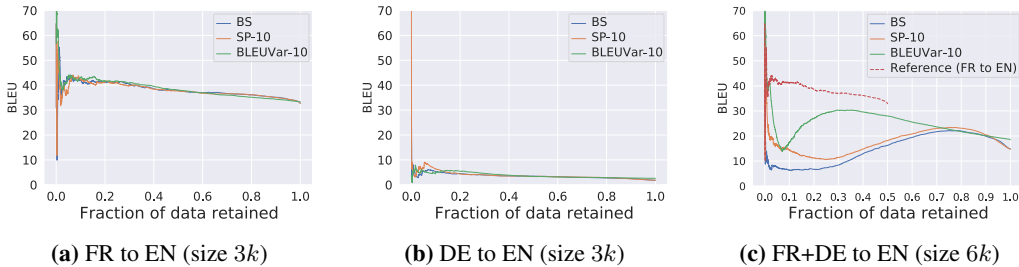
on the distribution of example uncertainty versus performance. We see that low data regimes lead to a low-entropy distribution with high uncertainty across the entire test set; as data availability is increased, uncertainty decreases, and average model performance increases for all rates of data retention.

## B ADDITIONAL OUT-OF-DISTRIBUTION EXPERIMENTS (FR+DE TO EN)

We have done similar experiments as section §4.1.2 on other language pair. Figure 9 here uses a similar experiment design as Figure 2 in §4.1.2. Instead of testing DE and NL on model trained with DE-EN task, here the model is trained with FR (French) to EN (English) task and tests on FR to EN (in-distribution), DE to EN (OOD) and FR+DE to EN test sets. The pair FR and EN has much less overlap in vocabulary than DE and NL.

The BLEU on the full combined test set (see Figure 9(c)) is the best the models can do; then BLEUVar rejects German sentences until it has mostly French sentences left and it has peak performance (data retained=0.3), after which it is forced to reject French sentences as well. Note that performance goes down for small retain rate because there is a small number of DE (OOD) data erroneously being assigned with high confidence, which our metric captures well.

This result is similar to the DE+NL experiments in §4.1.2, with BLEUVar outperforms the rest by a large margin in the mixed test set (see Figure 9(c) and Table 9(c)). Further, the BLEU scores of OOD test is roughly flat compared to the amount of data retained (see Figure 9(b)).



**Figure 9:** Uncertainty measure comparisons using the in-distribution FR-EN test set (a), out-of-distribution DE-EN test set (b) and the combined FR+DE to EN test set (c). The *Reference* line in (c) corresponds to the BS plot from (a), which only has 3k test data. Therefore it only reaches the fraction 0.5 in this graph. The model was trained for FR to EN task with the WMT 2014 English-French training set (size 36m) using 350k steps.

**Table 9:** AUC for all plots in Figure 9

	BS	SP-10	BLEUVar-10
(a) FR to EN	38.18	38.02	<b>38.53</b>
(b) DE to EN	3.56	3.78	<b>3.82</b>
(c) FR+DE to EN	14.70	17.67	<b>25.10</b>

## C RELATED WORK

Our work might look similar to quality estimation (QE) task in MT (Specia et al., 2010; Blatz et al., 2004), but the problem of QE is fairly different to what we do in this paper. QE assumes the existence of a fixed translation system (e.g., an in-house encoder-decoder attention-based NMT system, as in WMT19’s shared task in Quality Estimation). The QE models then have to determine the quality of the system’s output. In contrast, we look at the problem of “introspection” where the system has to decide the confidence (“quality”) of *its own* output. This confidence can then be used for selective classification where the model can reject some uncertain translation. Further, standard approaches in QE might assume access to privileged data (e.g., the NMT translations for the source sentences and their corresponding human post-edition, as in task 2 in WMT19’QE), which we do not require. In addition, most existing approaches for QE require additional model to be trained to estimate the translation quality of a MT model, while our method does not have such requirement. Therefore, our method is able to provide uncertainty estimate simply with the parallel corpus used for training the translation model without the need for additional data and training procedure.

The closest to our paper is task 3 in WMT19’QE: a metric to score sentences is sought, which must correlate to human judgement. We would like to stress that a system’s confidence in its own prediction does not have to be correlated to human judgement. Indeed, we demonstrate this in Appendix A.2 where a model can indicate that it does not have enough training data, and requires additional data to

increase its confidence (the model’s subjective view of its uncertainty does not have to correlate with empirical mistakes - Bayesian epistemology (Zalta et al., 1995)).

In addition, QE tasks are mainly focus on estimating the in-distribution translation quality, since the test sets are in the same domain as the training sets provided by WMT QE tasks (e.g. both in the IT domain for English-German WMT18,19). In contrast, the goal for our uncertainty estimate is to identify the out-of-distribution translations, rather than estimating the quality of in-distribution translation. Therefore, our tasks are fundamentally different to QE.

There have been some prior attempts at investigating the similar problem as ours. In particular, Kumar & Sarawagi (2019) investigated the calibration of various NMT models at the token level. Kumar & Sarawagi found that many models are ill-calibrated at the *token level*, leading to the resulting probability distribution over the vocabulary used during decoding is not a good reference for model uncertainty. To correct for this, Kumar & Sarawagi design a recalibration strategy that applies an adaptive temperature to the logits, determined by the token identities, attention entropies, and other relevant components. Desai & Durrett (2020) looked into the calibration of pre-trained Transformers, and discovered that pre-trained Transformers are well calibrated for in-distribution data but ill-calibrated for out-of-distribution data. Such observations on NMT calibration further motivate us to design better uncertainty measures for NMT models.

Another study of uncertainty in NMT models comes from Ott et al. (2018); they found models tend to have overly high uncertainty in their output distribution over sequences. Note that both do not consider *epistemic* uncertainty, not OOD settings. There are some work consider *epistemic* uncertainty (Fomicheva et al., 2020; Wang et al., 2019) and propose MC Dropout-based measures similar to our **Sequence Probability (SP)**. Our work explores this direction and offers a new uncertainty estimation technique (i.e.BLEUVar) that empirically out-performs existing methods by a significant margin.

## D TRANSLATION SAMPLES

### D.1 IN-DISTRIBUTION CERTAIN SAMPLES

**Table 10:** (Low uncertainty) In-distribution DE source sentence from the experiment in Figure 2(a).

<b>Source sentence (DE) :</b>	
Nevada hat bereits ein Pilotprojekt abgeschlossen.	
<b>Reference translation (EN) :</b> (only used to compute “BLEU to reference”)	
Nevada has already completed a pilot.	
<b>Model predictive-mean translation (EN) :</b> (averaging over predictive probabilities during decoding)	
Nevada has already completed a pilot project.	
<b>Translation “BLEU to reference” :</b>   70.7	
<b>Translation uncertainty :</b>   0	
<b>Translations sampled from the model:</b> (5 samples from predictive probabilities during decoding)	
1	Nevada has already completed a pilot project.
2	Nevada has already completed a pilot project.
3	Nevada has already completed a pilot project.
4	Nevada has already completed a pilot project.
5	Nevada has already completed a pilot project.

### D.2 IN-DISTRIBUTION UNCERTAIN SAMPLES

**Table 11:** (High uncertainty) In-distribution DE source sentence from the experiment in Figure 2(a).

<b>Source sentence (DE) :</b>	
Im Grunde genommen sind vegane Gerichte für alle da.	
<b>Reference translation (EN) :</b> (only used to compute “BLEU to reference”)	
Essentially, vegan dishes are for everyone.	
<b>Model predictive-mean translation (EN) :</b> (averaging over predictive probabilities during decoding)	
Basically vegan dishes are there for everyone.	
<b>Translation “BLEU to reference” :</b>   34.5	
<b>Translation uncertainty :</b>   3122	
<b>Translations sampled from the model:</b> (5 samples from predictive probabilities during decoding)	
1	Basically vegan dishes are for everyone.
2	Basically, vegan dishes are there for everyone.
3	Essentially, vegan dishes are available for everyone.
4	Basically, vegane dishes are there for all.
5	Basically vegan dishes are there for everyone.



