Non-Exchangeable Conformal Language Generation with Nearest Neighbors

Anonymous ACL submission

Abstract

Quantifying uncertainty in automatically generated text is important for letting humans check potential hallucinations and making systems more reliable. Conformal prediction is an attractive framework to provide predictions imbued with statistical guarantees, however, its application to text generation is challenging since any i.i.d. assumptions are not realistic. In this paper, we bridge this gap by leveraging recent results on *non-exchangeable* conformal prediction, which still ensures bounds on coverage. The result is a novel extension of the conformal prediction framework to generation based on nearest neighbors. Our method can be used post-hoc for an arbitrary model without extra training and supplies token-level, calibrated prediction sets equipped with statistical guarantees. Experiments in machine translation and language modeling show encouraging results in word coverage and generation quality.

1 Introduction

Natural language generation (NLG) is a multi-faceted field spanning applications such as machine translation (MT), language modeling (LM), summarization, question answering and dialogue generation. Owing to the recent success of large language models (LLMs) such as GPT-4 (OpenAI, 2023), BLOOM (Scao et al., 2022) or LLaMA (Touvron et al., 2023), natural language modeling with stochastic decoding (sampling) is increasingly used as an interface with end users. While sampling allows for more fluent and varied text, few methods exist to evaluate the reliability of generated text and adequacy of the underlying sampling method. This is particularly relevant for generation scenarios where pre-trained models are applied to new data with potentially different distribution to the training data, increasing the risk of generating erroneous, misleading, and potentially harmful text (Ji et al., 2023; Guerreiro et al., 2023; Pan et al., 2023; Alkaissi and McFarlane, 2023; Azamfirei et al., 2023).

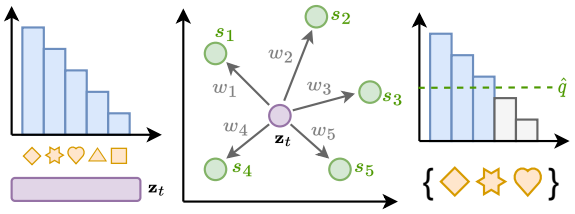


Figure 1: Schematic representation of our approach. A decoder hidden representation $\mathbf{z}_t$ is used during inference to retrieve the nearest neighbors and their non-conformity scores s_k. Their relevance is determined by using their distance to compute weights w_k, resulting in the quantile $\hat{q}$ that forms conformal prediction sets.

Conformal prediction (Vovk et al., 2005; Papadopoulos et al., 2002; Angelopoulos and Bates, 2021) has recently gained popularity by providing calibrated prediction sets that are imbued with statistical guarantees about containing the correct solution. Nevertheless, applying conformal prediction to NLG is not trivial and comes with a major obstacle: The conditional generation process breaks the independence and identical distribution (i.i.d.) assumption underlying conformal prediction techniques. We tackle this problem by drawing inspiration from recent advances in nearest neighbor language modeling (Khandelwal et al., 2020b; He et al., 2021a; Xu et al., 2023) and machine translation (Khandelwal et al., 2020a; Zheng et al., 2021; Meng et al., 2022; Martins et al., 2022). This way, we are able to dynamically generate calibration sets during inference that are able to maintain statistical guarantees. We schematically illustrate our method in Figure 1.

Contributions. We present a general-purpose extension of the conformal framework to NLG by

tackling the problems above. Our contributions are as follows: ① To the best of our knowledge, we are the first to present a novel technique based on *non-exchangeable* conformal prediction and to apply it to language generation to produce calibrated prediction sets. ② We validate the effectiveness of the method in a Language Modeling and Machine Translation context, evaluating the coverage of the calibrated prediction sets and showing that our method is on par with or even outperforms other sampling-based techniques in terms of generation quality, all while maintaining tighter prediction sets and better coverage. ③ We finally demonstrate that these properties are also maintained under distributional shift induced by corrupting the model's latent representations. ④ We publish all the code for this project in an open-source repository.[1]

2 Related Work

Conformal Prediction. Conformal prediction is a line of work that has recently regained interest in machine learning by producing prediction sets with certain statistical guarantees about containing the correct prediction (Vovk et al., 2005; Papadopoulos et al., 2002; Angelopoulos and Bates, 2021). As the size of prediction sets is calibrated to fulfill these guarantees, one can also see the size of the prediction set itself as a proxy of the uncertainty of a model—the larger the set, the more possible predictions have to be included in order to maintain the coverage guarantee. Conformal prediction has already found diverse applications in NLP for classification (Maltoudoglou et al., 2020; Fisch et al., 2021; Schuster et al., 2021; Fisch et al., 2022; Choubey et al., 2022; Kumar et al., 2023) and sequence labeling problems (Dey et al., 2021), as well as quality estimation (Giovannotti, 2023; Zerva and Martins, 2023). Unfortunately, generation problems are challenging due to their sequential nature and constant breaking of the i.i.d. assumption, so existing works operate on the sequence-level instead (Quach et al., 2023; Ren et al., 2023; Deutschmann et al., 2023). Conformal procedures for time-series (Xu and Xie, 2021; Lin et al., 2022b; Oliveira et al., 2022; Zaffran et al., 2022) and general non-i.i.d. data (Tibshirani et al., 2019; Barber et al., 2023; Guan, 2023; Farinhas et al., 2023) have been proposed in the literature. The most related work to ours is given by Ravfogel et al. (2023), who apply the standard conformal prediction setup to NLG, arguing that Markov chains are a type of β-mixing processes, for which Oliveira et al. (2022) showed coverage to degrade by an only negligible amount. However, Ravfogel et al. do not investigate this claim empirically, and furthermore do not find any benefits when generating sequences. In another related work, Quach et al. (2023) propose an approach that is specifically tailored toward language modeling. However, their prediction sets contain entire sequences instead of single tokens. In contrast, our token-level prediction sets are useful for constraining the options during generation and their widths can represent model uncertainty.

Uncertainty in NLP. Modeling uncertainty in NLP has already been studied in classification (Van Landeghem et al., 2022; Ulmer et al., 2022a; Holm et al., 2022) and regression settings (Beck et al., 2016; Glushkova et al., 2021; Zerva et al., 2022). However, NLG proves more challenging due to it non-i.i.d. and combinatorial nature. Some works have proposed Bayesian Deep Learning methods for NLG: Xiao et al. (2020) use Monte Carlo Dropout (Gal and Ghahramani, 2016) to produce multiple generations for the same input and measure their pair-wise BLEU scores. Malinin and Gales (2021) define extensions of mutual information for structured prediction. Other existing approaches try to account for the paraphrastic nature of language by modeling the entropy over meaning classes (Kuhn et al., 2023), investigate the use of linguistic markers to indicate uncertainty (Zhou et al., 2023) or ask the model directly for its confidence (Lin et al., 2022a; Kadavath et al., 2022). Baan et al. (2023) provide an extensive overview of the theory and current state of the field.

3 Background

Conformal Prediction. Conformal prediction is an attractive method for uncertainty quantification due to its statistical coverage guarantees (Vovk et al., 2005; Papadopoulos et al., 2002; Angelopoulos and Bates, 2021). Given some predictor, a held-out calibration set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, and a pre-defined miscoverage level α (e.g., 0.1), the calibration set is used to obtain *prediction sets* $\mathcal{C}(\mathbf{x}^*)$ for a new test point $\mathbf{x}^*$ satisfying

$$p\Big(y^* \in \mathcal{C}(\mathbf{x}^*)\Big) \geq 1 - \alpha, \qquad (1)$$

that is, the probability of the prediction set $\mathcal{C}(\mathbf{x}^*)$ containing the correct label y^* is *at least* $1-\alpha$. This

[1] Made available upon acceptance.

is achieved by the following recipe: Firstly, one has to define a *non-conformity score*, that provides an estimate of the distance of the test point to the rest of the data, i.e., a proxy for the uncertainty over the test point predictions. In this context, the score can be as simple as $s_i = 1 - p_{\boldsymbol{\theta}}(y|\,\mathbf{x})$, i.e. one minus the softmax probability of the true class, which will be higher when the model is wrong or less confident. Next, we define $\hat{q}$ as the $\lceil (N+1)(1-\alpha)/N \rceil$-th quantile of the non-conformity scores. Then, when we make a new prediction for a test point $\mathbf{x}^*$, we can create prediction sets defined as

$$\mathcal{C}(\mathbf{x}^*) = \Big\{ y \;\Big|\; p_{\boldsymbol{\theta}}(y|\,\mathbf{x}^*) \geq 1 - \hat{q} \Big\}, \qquad (2)$$

which is guaranteed to fulfil the coverage requirement in Equation (1) for i.i.d. data (Vovk et al., 2005; Angelopoulos and Bates, 2021).

Non-exchangeable Conformal Prediction. Barber et al. (2023) address a major shortcoming in the method above: When a test point and the calibration data are not i.i.d.,[2] the distributional drift causes any previously found $\hat{q}$ to be miscalibrated, and thus the intended coverage can no longer be guaranteed. However, we can still perform conformal prediction by assigning a weight $w_i \in [0, 1]$ to every calibration data point, reflecting its relevance—i.e. assigning lower weights to points far away from the test distribution. Then, by normalizing the weights with $\tilde{w}_i = w_i / (1 + \sum_{i=1}^{N} w_i)$, we define the quantile as

$$\hat{q} = \inf \Big\{ q \;\Big|\; \sum_{i=1}^{N} \tilde{w}_i \mathbf{1}\{s_i \leq q\} \geq 1 - \alpha \Big\}, \qquad (3)$$

with $\mathbf{1}\{\cdot\}$ denoting the indicator function. The construction of the prediction sets then follows the same steps as before. Most notably, the coverage guarantee in Equation (1) now changes to

$$p\Big(y^* \in \mathcal{C}(\mathbf{x}^*)\Big) \geq 1 - \alpha - \sum_{i=1}^{N} \tilde{w}_i \varepsilon_i, \qquad (4)$$

with an extra term including the *total variation distance* between the distribution of a calibration and a test point, $\varepsilon_i = d_{\mathrm{TV}}\big((\mathbf{x}_i, y_i), (\mathbf{x}^*, y^*)\big)$.[3] Unfortunately, this term is hard to estimate or bound, nevertheless, the selection of appropriate weights that can capture the relevance of calibration points to the test set should moderate both the impact of the distant data points on the estimation of the prediction set and the impact of d_{TV} on the coverage bound. In other words, for large d_{TV} values we expect to have smaller weights, that allow us to achieve coverage close to the desired values. We show in our experiments that the loss of coverage when using nearest neighbor weights is limited and revisit the practical implications in Section 5.

3.1 Method: Non-exchangeable Conformal Prediction through Nearest Neighbors

We now present a novel method to apply conformal prediction in NLG by synthesizing the non-exchangeable approach of Barber et al. (2023) with k-NN search-augmented neural models (Khandelwal et al., 2020a,b). The related approach by Ravfogel et al. (2023) calibrates prediction sets within bins of similar entropies using the non-exchangeable procedure described in Section 3. However, this implies that we would use semantically unrelated (sub-)sequences to calibrate the model—in fact, we show experimentally that this approach obtains generally trivial coverage by producing extremely wide prediction sets. Instead, we propose to perform a *dynamic* calibration step during model inference, only considering the most relevant data points from the calibration set. We do this in the following way: Given a dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}$ of sequences $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_S^{(i)})$ and corresponding references consisting of gold tokens $y^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})$, we extract the model's decoder activations $\mathbf{z}_t^{(i)} \in \mathbb{R}^d$ and conformity scores $s_t^{(i)}$.[4] We save those in a datastore allowing for fast and efficient nearest neighbor search using FAISS (Johnson et al., 2019). In the inference phase, during every decoding step, we then use the decoder hidden state $\mathbf{z}_t^*$ to query the data store for the K nearest neighbors and their conformity scores and record their distances. We use the squared l_2 distance to compute the weight w_k for a neighbor as

$$w_k = \exp\Big(- \big|\big|\, \mathbf{z}_t^* - \mathbf{z}_k \big|\big|_2^2 \,/\, \tau \Big), \qquad (5)$$

where τ corresponds to a temperature hyperparameter. Overall, this formulation is equivalent to a

[2] In fact, the coverage guarantee applies to the case where the data is *exchangeable*, a weaker requirement than i.i.d. Specifically, a series of random variables is exchangeable if their joint distribution is unaffected by a change of their order.

[3] In this expression, $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}^*, y^*)$ denote random variables and the total variation distance is between the two underlying distributions. See Barber et al. (2023) for details.

[4] In this phase, we do not let the model generate freely, but feed it the gold prefix during the decoding process to make sure that conformity scores can be computed correctly.

radial basis function kernel with scale parameter τ. Finally, we use the weights to compute the quantile $\hat{q}$ as in Equation (3). The entire algorithm is given in Appendix A.4.

Adaptive Prediction Sets. The efficacy of conformal prediction hinges on the choice of non-conformity score, with the simple non-conformity score $s_i = 1 - p_{\boldsymbol{\theta}}(y_t | \mathbf{x}, y_{<t})$ known to undercover hard and overcover easy subpopulations of the data. Due to the diverse nature of language, we therefore opt for *adaptive prediction sets* (Angelopoulos et al., 2021a; Romano et al., 2020). Adaptive prediction sets redefine the non-conformity score as the cumulative probability over classes necessary to reach the correct class. More formally, let π be a permutation function mapping all possible output tokens $\{1, \ldots, C\}$ to the indices of a permuted version of the set, for which tokens are sorted by their probability under the model, descendingly. We define the non-conformity score as

$$s_i = \sum_{j=1}^{\pi(y_t)} p_{\boldsymbol{\theta}}\big(\pi^{-1}(j) \big| \mathbf{x}, y_{<t}\big). \tag{6}$$

Since we only include the cumulative mass up until the gold label, the summation stops at $\pi(y)$. The prediction sets are then defined as

$$\mathcal{C}(\mathbf{x}^*, y^*_{<t}) = \Big\{\pi^{-1}(1), \ldots, \pi^{-1}(\hat{c})\Big\}, \tag{7}$$

with $\hat{c} = \sup\{c' \mid \sum_{j=1}^{c'} p_{\boldsymbol{\theta}}(\pi^{-1}(j) \mid \mathbf{x}^*, y^*_{<t}) < \hat{q}\} + 1$. Intuitively, this means that we included all classes whose cumulative probability (after sorting descendingly) does not surpass $\hat{q}$, adding one extra class to avoid empty sets. Compared to the simple conformity score, this produces wider predictions sets for hard inputs, encompassing more potentially plausible continuations in a language context.

4 Experiments

In the following sections, we conduct experiments in both language modeling and machine translation. For machine translation we opt for the 400 million and 1.2 billion parameter versions of the M2M100 model (Fan et al., 2021) on the WMT-2022 shared task datasets for German to English and Japanese to English (Kocmi et al., 2022). For Language Modelling, we use the 350 million and 1.3 billion parameter versions of the OPT model (Zhang et al., 2022) and replicate the setup by Ravfogel et al. (2023): We calibrate our model on 10000 sentences from a 2022 English Wikipedia dump (Foundation, 2022) and test coverage and generation on 1000 sentences from OpenWebText (Gokaslan et al., 2019).[5] All models are used in a zero-shot setup *without extra training or finetuning*. For the datastore, we use the implementation by FAISS library (Johnson et al., 2019), computing 2048 clusters in total and probing 32 clusters per query. We also summarize the environmental impact of our experiments in Appendix A.5.

4.1 Evaluating Coverage

First of all, we demonstrate that the retrieved information from the data store enables us to successfully apply the proposed method. While it is not possible to measure coverage in a free generation setting (see next section), we can assess whether the correct class is contained in the prediction set if we feed the actual reference tokens into the decoder and check whether we include the true continuation.[6] For our MT task, this is reminiscent of an interactive translation prediction setup (Knowles and Koehn, 2016; Peris et al., 2017; Knowles et al., 2019), where we would like to suggest possible continuations to a translator, suggesting the next word from a set of words that (a) contains plausible options and (b) is limited in size, in order to restrict the complexity for the end user. Before we run our experiments, we need to determine τ, which we tune on the calibration set using a stochastic hill-climbing procedure described in Appendix A.1. We compare our *non-exchangeable conformal nucleus sampling* (*Non-Ex. CS*) with nucleus sampling (Holtzman et al., 2020) and conformal nucleus sampling (*Conf. Sampl.*; Ravfogel et al., 2023), using 10 entropy bins and corresponding $\hat{q}$ values.

Evaluation. We evaluate by measuring the total coverage using different distance metrics, namely, squared l_2 distance, normalized inner product, and cosine similarity (see Tables 1 and 2),[7] as well as binning predictions by set size and then measuring the per-bin coverage in Figure 2 (more results given in Appendix A.2). We also summarize the plots in

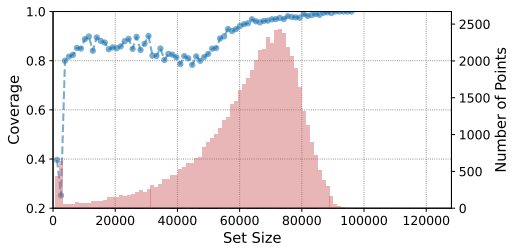
[5] Data obtained through the Hugging Face `datasets` package (Lhoest et al., 2021): https://huggingface.co/datasets/wikipedia and https://huggingface.co/datasets/stas/openwebtext-10k.

[6] We emphasize that access to gold tokens is not required by our method and only done here to measure the actual coverage.

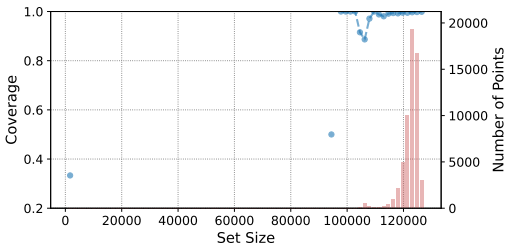
[7] For inner product and cosine similarity, we follow the same form as Equation (5), omitting the minus. We normalize the inner product by the square root of the latent dimension.

Table 1: Coverage results for the de → en and ja → en MT tasks. We report the best found temperature τ while keeping the confidence level α and number of neighbors $k = 100$ fixed. We also show the coverage percentage along with the avg. prediction set size as a proportion of the entire vocabulary (∅ WIDTH) as well as ECG and SSC. Tested distance metrics are inner product (IP), (squared) l_2 distance, and cosine similarity (cos).

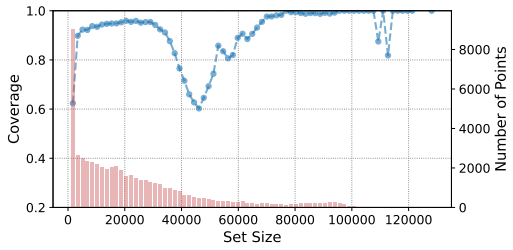
			de → en					ja → en				
	Method	Dist.	τ	% COVERAGE	∅ WIDTH ↓	SCC ↑	ECG ↓	τ	% COVERAGE	∅ WIDTH ↓	SCC ↑	ECG ↓
$M2M100_{(400M)}$	Nucleus Sampling	-	-	0.9207	0.48	0.25	0.00	-	0.9261	0.54	0.41	0.02
	Conf. Sampling	-	-	0.9951	0.94	0.33	0.03	-	0.9950	0.96	0.14	0.00
	Non-Ex. CS	IP	3.93	0.8251	0.16	0.63	0.26	11.90	0.8815	0.24	0.67	0.03
		l_2	512.14	0.8334	0.17	0.60	0.06	419.91	0.8468	0.18	0.61	0.05
		cos	2.54	0.8371	0.17	0.63	0.06	3.53	0.8540	0.17	0.62	0.04
$M2M100_{(1.2B)}$	Nucleus Sampling	-	-	0.8339	0.38	0.00	0.08	-	0.7962	0.42	0.03	0.10
	Conf. Sampling	-	-	0.9993	0.99	0.34	0.00	-	0.9998	0.99	0.60	0.00
	Non-Ex. CS	IP	15.79	0.8861	0.25	0.71	0.03	10.45	0.9129	0.38	0.72	0.00
		l_2	1123.45	0.8874	0.25	0.72	0.03	605.97	0.8896	0.30	0.76	0.01
		cos	3.21	0.8858	0.25	0.72	0.03	1.48	0.8897	0.30	0.75	0.01



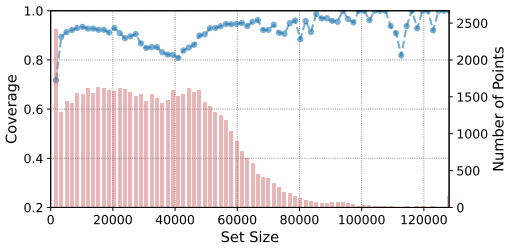
(a) Nucleus Sampling on de → en.



(b) Conformal Nucleus Sampling on de → en.



(c) Non-Ex. Conformal Sampling on de → en.



(d) Non-Ex. CS on de → en with $M2M100_{(1.2B)}$.

Figure 2: Conditional coverage for the M2M100 on de → en with the small 418M model (Figures 2a to 2c) and using the bigger 1.2B model (Figure 2d). We aggregate predictions by set size using 75 equally-spaced bins in total. The blue curve shows the conditional coverage per bin, whereas red bars show the number of binned predictions.

Figure 2 via the *Expected Coverage Gap* (ECG)[8] that we define as

$$\text{ECG} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{N} \max\Big(1-\alpha-\text{Coverage}\big(\mathcal{B}_b\big), 0\Big), \tag{8}$$

where $\mathcal{B}_b$ denotes a single bin and N the total number of considered predictions in the dataset.[9] In our experiments, we use 75 bins in total. The same bins are used to also evaluate the *Size-Stratified Coverage metric* (SSC) proposed by Angelopoulos et al. (2021b) to assess the balance of coverage across set sizes, with a well-calibrated method resulting in a SCC close to the desired coverage $1 - \alpha$:

$$\text{SCC} = \min_{b\in\{1,\dots,B\}} \text{Coverage}\big(\mathcal{B}_b\big) \tag{9}$$

[8] This measure is inspired by the expected calibration error (Guo et al., 2017), but measuring the coverage against a constant target value $1 - \alpha$. Since Conformal Prediction provides a lower bound, overcoverage is not penalized.

[9] Since conformal prediction produces a *lower* bound on the coverage, we do not include overcoverage in Equation (8).

Table 2: Coverage results for the LM task. We report the best found temperature τ while keeping the confidence level α and number of neighbors $k = 100$ fixed. We also show the coverage percentage along with the avg. prediction set size as a proportion of the entire vocabulary (∅ WIDTH) as well as the ECG and SSC metrics. Tested distance metrics are inner product (IP), (squared) l_2 distance and cos. similarity (cos).

			OpenWebText				
	Method	Dist.	τ	% Cov.	∅ Width ↓	Scc ↑	Ecg ↓
$OPT_{(350M)}$	Nucl. Sampl.	-	-	0.8913	0.05	0.71	0.01
	Conf. Sampl.	-	-	0.9913	0.90	0.91	0.00
	Non-Ex. CS	IP	4.99	0.9352	0.19	0.80	0.0
		l_2	0.31×10^4	0.9425	0.17	0.80	0.0
		cos	4.98	0.9370	0.15	0.83	0.0
$OPT_{(1.3B)}$	Nucl. Sampl.	-	-	0.8952	0.05	0.00	0.01
	Conf. Sampl.	-	-	0.9905	0.88	0.95	0.0
	Non-Ex. CS	IP	0.48	0.9689	0.59	0.84	0.0
		l_2	1.55×10^4	0.9539	0.20	0.83	0.0
		cos	0.11	0.9512	0.20	0.875	0.0

We present some additional experiments where we assess the impact of key hyperparameters in Appendix A.3.

Results. We found our method to miss the desired coverage of 90% for MT by 8% or less. Beyond the reported values, we were not able to further increase coverage by varying the temperature parameter without avoiding trivial coverage (i.e., defaulting to very large set sizes), which is likely due to the impossible-to-estimate coverage in Equation (4). Most notably, our method was able to achieve better SCC scores while maintaining considerably smaller prediction sets than the baselines on average. The reason for this is illustrated in Figure 2: while standard nucleus sampling produces some prediction sets that are small, the total coverage seems to mostly be achieved by creating prediction sets between 60k–80k tokens. The behavior of conformal nucleus sampling by Ravfogel et al. (2023) is even more extreme in this regard, while our method focuses on producing smaller prediction sets, with the frequency of larger set sizes decreasing gracefully. In Figure 2d, we can see that the larger M2M100 models also tend to produce larger prediction sets, but still noticeably smaller than the baselines. Importantly, for both M2M100 models, even very small prediction sets (size ≤ 1000) achieve non-trivial coverage, unlike the baseline methods. For LM, we always found the model to slightly *over*cover. This does not contradict the desired lower bound on the coverage in Equation (4) and suggests a more negligible distributional drift. While nucleus sampling produces the smallest average prediction sets, we can see that based on the SCC values some strata remain undercovered. Instead, our method is able to strike a balance between stratified coverage and prediction set size. With respect to distance measures, we find that the difference between them is minimal, indicating that the quality largely depends on the retrieved local neighborhood of the decoder encoding and that finding the right temperature can help to tune the models to approximate the desired coverage. Now we would like to find out whether this neighborhood retrieval mechanism can prove to be robust under distributional shift as well.

4.2 Coverage Under Shift

To demonstrate how the retrieval of nearest neighbors can help to maintain coverage under distributional shift, we add Gaussian noise of increasing variance—and therefore intensity—to the last decoder hidden embeddings (for MT) and the input embeddings (LM). This way, we are able to simulate distributional drift while still keeping the original sequence of input tokens intact, allowing us to measure the actual coverage. We show the achieved coverage along with the average set size (as a percentage of the total vocabulary) and the average quantile $\hat{q}$ in Figure 3. We can see that the conformal sampling method deteriorates into returning the full vocabulary as a prediction set. Thus it behaves similarly to simple sampling as indicated by the $\hat{q}$ values being close to 1. Nucleus sampling provides smaller prediction sets compared to conformal sampling, but they seem invariant to noise. As such, the method is not robust to noise injection in the open text generation task, and the obtained coverage deteriorates with noise variance ≥ 0.025. Instead, the use of nearest neighbors allows for the estimation of prediction sets that are small but amenable to increase, such that the obtained coverage remains close to the desired one. We can specifically observe that the prediction set size increases considerably to mitigate the injected noise in the open-text generation case.

Neighbor Retrieval. We further analyze how the retrieval enables this flexibility by relating it to the entropy of the output distribution of the 400M parameters M2M100 on German to English. Intuitively, the baseline methods, faced by high-entropy output distributions, need to produce wide predic-

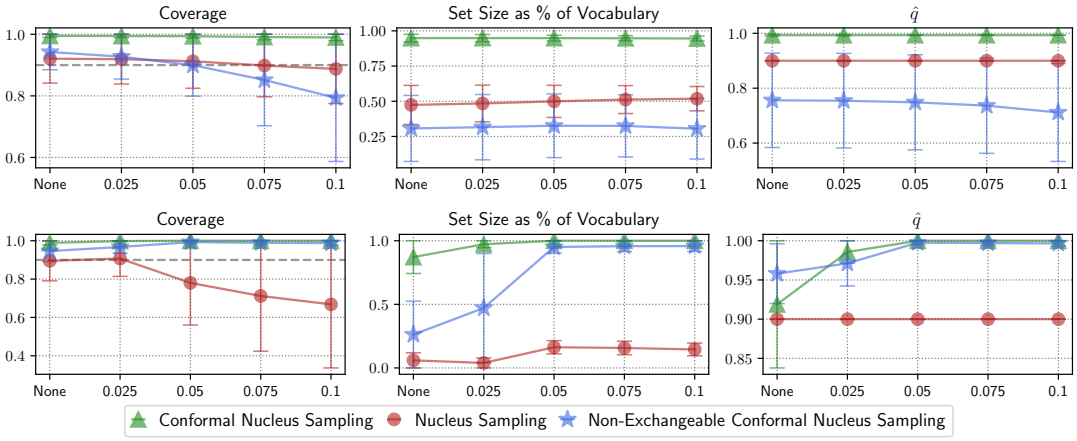


Figure 3: Coverage, average set size and $\hat{q}$ based on the noise level on the de $\rightarrow$ en MT task (top) and open text generation task (bottom). Error bars show one standard deviation.

Table 3: Average entropy of 400M M2M100 model on de $\rightarrow$ en per noise level as well as the Spearman's ρ correlation coefficients between the predictive entropy and the prediction set size of the different methods. All results are significant with $p < 0.0001$.

	Noise level				
	None	0.025	0.05	0.075	0.1
∅ Entropy	8.46	8.71	9.20	9.71	10.08
Nucl. Sampl. (ρ)	0.87	0.86	0.84	0.82	0.81
Conf. Sampl. (ρ)	0.60	0.60	0.60	0.57	0.55
Non-Ex. CS (ρ)	−0.14	−0.18	−0.27	−0.37	−0.45

tion sets in order to maintain coverage. In fact, we report such results by correlating entropy levels and prediction set sizes using Spearman's ρ in Table 3, showing strong positive correlations. Our method in contrast shows consistently an *anti*correlation between these two quantities, enabled by decoupling the creation of prediction sets from statistics of the output distribution to instead considering the non-conformity scores of similar subsequences. The fact that the prediction set size is not just dependent on the entropy of the predictions while maintaining coverage demonstrates the value of the nearest neighbors: In this way, model uncertainty becomes more flexible and is corroborated by evidence gained from similar inputs.

4.3 Generation Quality

Crucially, our method should not degrade and potentially even improve generation quality. Thus, we evaluate generation quality for the same tasks without supplying the gold prefix. For language modeling, we follow Ravfogel et al. (2023) and use the first 35 tokens from the original sentence as input. We compare against a set of generation strategies including top-k sampling (Fan et al., 2018; Holtzman et al., 2018; Radford et al., 2019), nucleus sampling and conformal nucleus sampling. We also test a variant of our method using constant weights $w_k = 1$ for retrieved neighbors (*Const. Weight CS*) to assess the impact of the weighted neighbor retrieval procedure. We further compare with beam search (Medress et al., 1977; Graves, 2012) with a softmax temperature of 0.1, and greedy decoding. Evaluation is performed using BLEU (Papineni et al., 2002), COMET-22 (Rei et al., 2020, 2022) and chrF (Popović, 2017) for MT as well as MAUVE (Pillutla et al., 2021) and BERTscore (Zhang et al., 2020) for text generation.[10]

Results. We show the results for the different methods in Table 4. We see that beam search outperforms all sampling methods for MT. This corroborates previous work by Shaham and Levy (2022) who argue that (nucleus) sampling methods, by pruning only the bottom percentile of the token distribution, introduce some degree of randomness that is beneficial for open text generation but may be less optimal for conditional language generation, where the desired output is constrained and exact matching generations are preferred (which is the case for MT). Among sampling methods, we find nucleus sampling and conformal sampling to perform similarly (being in agreement with the findings of Ravfogel et al., 2023) but are

[10] All metrics except for COMET were used through Hugging Face `evaluate`. MAUVE uses `gpt2` as a featurizer.

	Method	de → en BLEU ↑	de → en COMET ↑	de → en CHRF ↑	ja → en BLEU ↑	ja → en COMET ↑	ja → en CHRF ↑
$M2M100_{(400m)}$	Beam search	28.53	0.88	55.58	11.37	0.63	37.74
	Greedy	27.81	0.9	54.9	10.73	0.58	36.5
	Nucleus Sampling	27.63 ±0.03	0.89 ±0.01	54.80 ±0.07	10.61 ±0.15	0.59 ±0.01	36.52 ±0.19
	Top-k Sampling	27.63 ±0.03	0.89 ±0.01	54.79 ±0.07	10.61 ±0.15	0.59 ±0.01	36.52 ±0.19
	Conf. Sampling	27.63 ±0.03	0.89 ±0.01	54.80 ±0.07	10.61 ±0.15	0.59 ±0.01	36.52 ±0.19
	Const. Weight CS*	27.63 ±0.03	0.89 ±0.01	54.80 ±0.07	10.61 ±0.15	0.59 ±0.01	36.52 ±0.19
	Non-Ex. CS*	27.65 ±0.10	0.90 ±0.01	54.82 ±0.14	10.74 ±0.11	0.59 ±0.01	36.61 ±0.08
$M2M100_{(1.2B)}$	Beam search	30.89	0.9	56.8	13.76	0.63	40.43
	Greedy	29.52	0.9	55.67	12.94	0.6	39.91
	Nucleus Sampling	29.37 ±0.12	0.90 ±0.00	55.55 ±0.11	10.61 ±0.15	0.59 ±0.01	36.52 ±0.19
	Top-k Sampling	29.53 ±0.00	0.90 ±0.00	55.67 ±0.00	12.91 ±0.08	0.60 ±0.01	39.95 ±0.00
	Conf. Sampling	29.37 ±0.12	0.90 ±0.00	55.55 ±0.11	12.91 ±0.08	0.60 ±0.00	39.95 ±0.08
	Const. Weight CS*	29.37 ±0.12	0.90 ±0.00	55.55 ±0.11	12.91 ±0.08	0.60 ±0.01	39.95 ±0.08
	Non-Ex. CS*	29.37 ±0.12	0.90 ±0.00	55.55 ±0.11	12.91 ±0.08	0.60 ±0.01	39.95 ±0.08

(a) Generation results for the de → en and ja → en translation tasks.

	Method	OpenWebText MAUVE ↑	OpenWebText BERTScore F_1 ↑
$OPT_{(350M)}$	Beam search	0.12	0.79
	Greedy	0.02	0.79
	Nucleus Sampling	0.91 ±0.02	0.80 ±0.00
	Top-k Sampling	0.90 ±0.03	0.80 ±0.00
	Conf. Sampling	0.91 ±0.02	0.80 ±0.00
	Const. Weight CS*	0.91 ±0.02	0.80 ±0.00
	Non-Ex. CS*	0.92 ±0.01	0.80 ±0.00
$OPT_{(1.3B)}$	Beam search	0.17	0.80
	Greedy	0.05	0.79
	Nucleus Sampling	0.91 ±0.02	0.80 ±0.00
	Top-k Sampling	0.93 ±0.01	0.81 ±0.00
	Conf. Sampling	0.93 ±0.01	0.80 ±0.00
	Const. Weight CS*	0.91 ±0.02	0.80 ±0.00
	Non-Ex. CS*	0.92 ±0.01	0.81 ±0.00

(b) Results for the open text generation.

Table 4: Generation results for the two tasks. We report performance using 5 beams for beam-search, top-k sampling with $k = 10$, and nucleus sampling with $p = 0.9$. Conformal methods all use $\alpha = 0.1$, with non-exchangeable variants retrieving 100 neighbors. MT results for sampling use a softmax temperature of 0.1. Our methods are marked with *. Results using 5 different seeds that are stat. significant according to the ASO test (Del Barrio et al., 2018; Dror et al., 2019; Ulmer et al., 2022b) with a confidence level of 0.95 and threshold $\varepsilon_{\text{min}} \leq 0.3$ are underlined.

sometimes on par or even outperformed by our non-exchangeable conformal sampling for MT. For text generation, our method performs best for the smaller OPT model but is slightly beaten by conformal nucleus sampling in terms of MAUVE. When using constant weights, performance deteriorates to the conformal sampling setup, emphasizing the importance of not considering all conformity scores equally when computing $\hat{q}$, even though the effect seems to be less pronounced for larger models. This illustrates the benefit of creating flexible prediction sets that are adapted on token-basis, suggesting that both the latent space neighborhoods induced by the model as well as the conformity scores are informative.

5 Discussion

Our experiments have shown that despite the absence of i.i.d. data in NLG and the loss in coverage induced by using dynamic calibration sets, the resulting coverage is still close to the pre-specified desired level for both LM and MT. Additionally, even though the coverage gap predicted by the method of Barber et al. (2023) is infeasible to quantify for us, we did not observe any critical degradation in practice. Further, we demonstrated how sampling from these calibrated prediction sets performs similarly or better than other sampling methods. Even though our method is still outperformed by beam search in the MT setting, previous work such as minimum bayes risk decoding has shown how multiple samples can be re-ranked to produce better outputs (Kumar and Byrne, 2004; Eikema and Aziz, 2020; Freitag et al., 2023; Fernandes et al., 2022). Additionally, recent dialogue systems based on LLMs use sampling instead of beam search for generation. Since our prediction sets are more flexible and generally tighter, our results serve as a starting point for future work. For instance, our technique could be used with new non-conformity scores that do not consider token probabilities alone (e.g. Meister et al., 2023) or using prediction set widths as a proxy for model uncertainty (Angelopoulos et al., 2021a).

6 Conclusion

We successfully demonstrated the application of a non-exchangeable variant of conformal prediction to machine translation and language modeling with the help of k-NN retrieval. We showed our method to be able to maintain the desired coverage best across different dataset strata while keeping prediction sets smaller than other sampling methods. We validated our method to produce encouraging results for generation tasks. Lastly, we analyzed the behavior under distributional drift, showing how the k-NN retrieval maintains desirable properties for the estimated prediction sets.

Limitations

We highlight two main limitations of our work here: Potential issues arising from different kinds of dataset shift as well as efficiency concerns.

Distributional Drifts. Even though any loss of coverage due to the term quantifying distributional drift in Equation (4) was limited in our experiments (see Sections 4.1 and 4.2), this might not hold across all possible setups. As long as we cannot feasibly approximate the shift penalty, it is impossible to determine a priori whether the loss of coverage might prove to be detrimental, and would have to be checked in a similar way as in our experiments. Furthermore, we only consider shifts between the models' training distributions and test data distributions here, while many other, unconsidered kinds of shifts exist (Moreno-Torres et al., 2012; Hupkes et al., 2022).

Computational Efficiency. Even using optimized tools such as FAISS (Johnson et al., 2019), moving the conformal prediction calibration step to inference incurs additional computational cost during generation. Nevertheless, works such as He et al. (2021b); Martins et al. (2022) show that there are several ways to improve the efficiency of k-NN approaches, and we leave such explorations to future work.

Ethical Considerations

The main promise of conformal prediction lies in its correctness—i.e. producing prediction sets that contain the correct prediction and are thus reliable. In an application, this could potentially create a false sense of security. On the one hand, the conformal guarantee holds in expectation, and not necessarily on a per-sample basis. On the other hand, our experiments have demonstrated that coverage might also not hold when distributional shifts are at work or when looking at specific subpopulations. Therefore, any application should certify that coverage is maintained for potentially sensitive inputs.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. 2021a. Uncertainty sets for image classifiers using conformal prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. 2021b. Uncertainty sets for image classifiers using conformal prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

Daniel Beck, Lucia Specia, and Trevor Cohn. 2016. Exploring prediction uncertainty in machine translation quality estimation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 208–218, Berlin, Germany. Association for Computational Linguistics.

Prafulla Kumar Choubey, Yu Bai, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2022. Conformal predictor for improving zero-shot text classification efficiency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3034, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Nicolas Deutschmann, Marvin Alberts, and María Rodríguez Martínez. 2023. Conformal autoregressive generation: Beam search with coverage guarantees. *arXiv preprint arXiv:2309.03797*.

Neil Dey, Jing Ding, Jack Ferrell, Carolina Kapper, Maxwell Lovig, Emiliano Planchon, and Jonathan P Williams. 2021. Conformal prediction for text infilling and part-of-speech prediction. *arXiv preprint arXiv:2111.02592*.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*

2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2773–2785. Association for Computational Linguistics.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André FT Martins. 2023. Non-exchangeable conformal risk control. *arXiv preprint arXiv:2310.01262*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José Guilherme Camargo de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1396–1412. Association for Computational Linguistics.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2021. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pages 3329–3339. PMLR.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2022. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*, pages 6514–6532. PMLR.

Wikimedia Foundation. 2022. Wikimedia downloads.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation. *arXiv preprint arXiv:2305.09860*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Patrizio Giovannotti. 2023. Evaluating machine translation quality with conformal predictive distributions. *arXiv preprint arXiv:2306.01549*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Leying Guan. 2023. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50.

Nuno Miguel Guerreiro, Elena Voita, and André F. T. Martins. 2023. 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, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1059–1075. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021a. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5703–5714. Association for Computational Linguistics.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021b. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5703–5714. Association for Computational Linguistics.

Andreas Nugaard Holm, Dustin Wright, and Isabelle Augenstein. 2022. Revisiting softmax for uncertainty approximation in text classification. *arXiv preprint arXiv:2210.14037*.

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

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1638–1649. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2022. State-of-the-art generalisation research in nlp: a taxonomy and review. *arXiv preprint arXiv:2210.03050*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020a. Nearest neighbor machine translation. *arXiv preprint arXiv:2010.00710*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020b. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120.

Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. 2019. A user study of neural interactive translation prediction. *Machine Translation*, 33:135–154.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

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

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2022b. Conformal prediction intervals with temporal dependence. *arXiv preprint arXiv:2205.12940*.

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Conformal and Probabilistic Prediction and Applications*, pages 269–284. PMLR.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2022. Chunk-based nearest neighbor machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4228–4245. Association for Computational Linguistics.

Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea, et al. 1977. Speech understanding systems: Report of a steering committee. *Artificial Intelligence*, 9(3):307–316.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 555–565. Association for Computational Linguistics.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. 2022. Split conformal prediction for dependent data. *arXiv preprint arXiv:2203.15885*.

OpenAI. 2023. Gpt-4 technical report.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer.

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.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2023. Conformal language modeling.

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

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. *arXiv preprint arXiv:2305.02633*.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. 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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. 2021. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing.

Tal Schuster, Adam Fisch, Tommi S. Jaakkola, and Regina Barzilay. 2021. Consistent accelerated inference via confident adaptive transformers. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 4962–4979. Association for Computational Linguistics.

Uri Shaham and Omer Levy. 2022. What do you get when you cross beam search with nucleus sampling? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 38–45.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022a. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022b. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*.

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. 2022. Benchmarking scalable predictive uncertainty in text classification. *IEEE Access*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Tim Z Xiao, Aidan N Gomez, and Yarin Gal. 2020. Wat zei je? detecting out-of-distribution translations with variational transformers. *arXiv preprint arXiv:2006.08344*.

Chen Xu and Yao Xie. 2021. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR.

Frank F Xu, Uri Alon, and Graham Neubig. 2023. Why do nearest neighbor language models work? *arXiv preprint arXiv:2301.02828*.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. 2022. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022. Disentangling uncertainty in machine translation evaluation.

Chrysoula Zerva and André FT Martins. 2023. Conformalizing machine translation evaluation. *arXiv preprint arXiv:2306.06221*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor 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, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 368–374. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

A Experimental Appendix

In this appendix, we bundle more details about experiments and their results. Appendix A.1 details the procedure to determine the temperature in Equation (5). We present more results from the experiments in Section 4.1 in Appendix A.2.

We illustrate the overall algorithm in Appendix A.4 and estimate environmental impact of our work in Appendix A.5.

A.1 Temperature Search

In order to determine the temperature used in Equation (5) for the different distance metrics in Table 1, we adopt a variation of a simple hill-climbing procedure. Given user-defined bounds for the temperature search τ_{min} and τ_{max}, we sample an initial candidate $\tau_0 \sim \mathcal{U}[\tau_{\text{min}}, \tau_{\text{max}}]$, and then evaluate the coverage of the method given the candidate on the first 100 batches of the calibration dataset. The next candidate then is obtained via

$$\tau_{t+1} = \tau_t + \eta \cdot \varepsilon \cdot \text{sgn}\big(1 - \alpha - \text{Coverage}(\tau_t)\big); \quad \varepsilon \sim \mathcal{N}(0, \tau_{\text{max}} - \tau_{\text{min}}), \tag{10}$$

where η is a predefined step size (in our case 0.1) and Coverage(τ_t) the achieved coverage given a candidate τ_t. The final temperature is picked after a fixed number of steps ($t = 20$ in our work) based on the smallest difference between achieved and desired coverage.

Overall, we found useful search ranges to differ greatly between datasets, models, and distance metrics, as illustrated by the reported values in Table 1 and Table 2. In general, the stochastic hill-climbing could also be replaced by a grid search, even though we sometimes found the best temperature to be "hidden" in a very specific value range. It also has to be noted that temperature for the l_2 distance is the highest by far since FAISS returns *squared* l_2 distances by default.

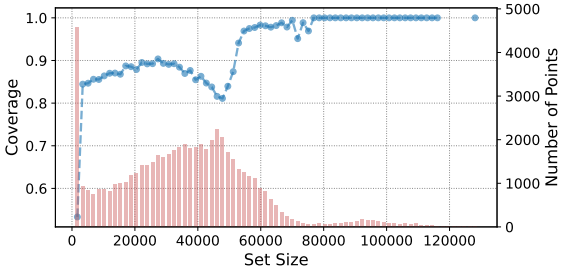
A.2 Additional Coverage Results

We show additional plots illustrating the coverage per set size-bins in Figure 4. We can see the counterparts for Figure 2 using the larger M2M100$_{(1.2B)}$ model in Figures 4a and 4b: Instead of leveling off like for the smaller model, most prediction set sizes are either in a very small range or in a size of a few ten thousand. In Figures 4c and 4d, we show similar plots for the two different OPT model sizes. Since in both cases, most prediction set sizes are rather small, we zoom in on the the sizes from 1 to 100. Here, we can observe a similar behavior to the smaller M2M100$_{(400m)}$, gradually leveling off. We do not show similar plots for other distance metrics as they show similar trends.

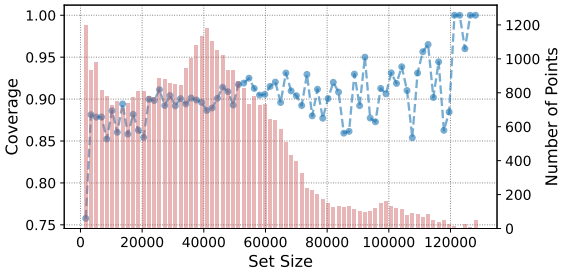
A.3 Impact of Coverage Threshold and Neighborhood Size Choice

In this section, we present experiments surrounding the two most pivotal parameters of our method: The desired confidence level α, as well as the number of neighbors.

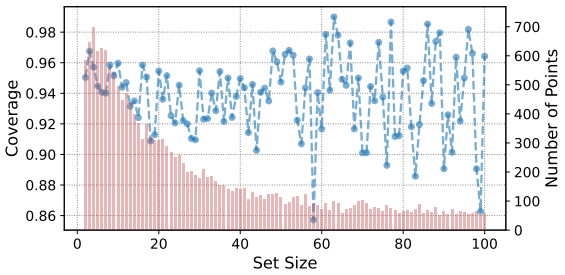
Coverage Threshold. In Table 5, we investigate the impact of different values on α on our evaluation metrics. We show that the increase in α does indeed produce the expected decrease in coverage, however with a certain degree of overcoverage for the de $\rightarrow$ en MT and the LM task. The loss in coverage always goes hand in hand with a decrease in the average prediction set width as well, as the model can allow itself to produce tighter prediction sets at the cost of higher miscoverage. As this also produces bin in which all contained instances are uncovered, this produces zero values for the SCC, while we cannot discern clearn trends for the ECG.



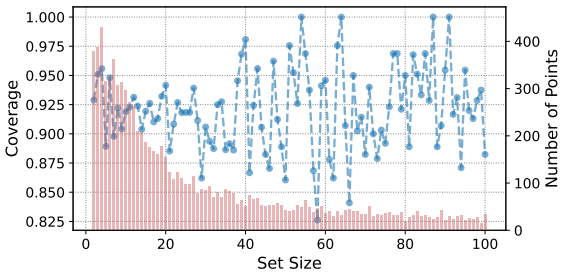
(a) Conditional coverage of M2M100$_{(1.2B)}$ for de $\rightarrow$ en.



(b) Conditional coverage of M2M100$_{(1.2B)}$ for ja $\rightarrow$ en.



(c) Conditional coverage for OPT$_{(350M)}$ on Language Modelling.



(d) Conditional coverage for OPT$_{(1.3B)}$ on Language Modelling.

Figure 4: Additional conditional coverage plots for the MT and LM dataset using our non-exchangeable conformal prediction method, aggregating predictions by prediction set size. The blue curve shows the conditional coverage per bin, whereas red bars show the number of predictions per bin. For Figures 4c and 4d, we zoom in on the prediction set sizes from 1 and 100.

Neighborhood Size. In Table 6, we vary the effect of the chosen neighborhood size (with 100

Table 5: Results for different values of α using different models and datasets.

	α	% Cov.	∅ Width ↓	Scc ↑	Ecg ↓
M2M100$_{(400M)}$ / de → en	0.1	0.9442	0.31	0.8702	0.0011
	0.2	0.8767	0.18	0.7906	8.63×10^{-5}
	0.3	0.7963	0.12	0	0.0016
	0.4	0.7058	0.09	0.1393	0.0082
	0.5	0.6081	0.07	0.2836	0.0055
	0.6	0.5017	0.06	0.1393	0.0082
	0.7	0.3896	0.05	0	0.0091
	0.8	0.2800	0.05	0	0.0090
	0.9	0.1762	0.04	0	0.0071
M2M100$_{(400M)}$ / ja → en	0.1	0.7453	0.15	0.3080	0.1511
	0.2	0.5579	0.07	0.2728	0.2446
	0.3	0.4277	0.04	0.2770	0.2779
	0.4	0.3438	0.03	0.1212	0.2438
	0.5	0.2749	0.03	0.0455	0.1883
	0.6	0.2175	0.02	0	0.1207
	0.7	0.1685	0.02	0	0.0560
	0.8	0.1309	0.01	0	0.0117
	0.9	0.0989	0.02	0	0.0099
OPT$_{(350M)}$ / OpenWebText	0.1	0.9460	0.26	0.8	1.85×10^{-5}
	0.2	0.8937	0.16	0.8	0
	0.3	0.8392	0.10	0.5	8.74×10^{-6}
	0.4	0.7782	0.08	0.6667	0
	0.5	0.7171	0.06	0	1.19×10^{-5}
	0.6	0.6559	0.06	0.6033	0
	0.7	0.5945	0.05	0	8.21×10^{-6}
	0.8	0.5349	0.05	0.4462	0
	0.9	0.4757	0.05	0.3580	0

Table 6: Results for different neighborhood sizes K using different models and datasets.

	K	% Cov.	∅ Width ↓	Scc ↑	Ecg ↓
M2M100$_{(400M)}$ / de → en	10	0.9923	0.39	0.9728	0
	25	0.9563	0.37	0.8877	0.0011
	50	0.9504	0.32	0.8870	0.0006
	75	0.9444	0.32	0.8641	0.0014
	100	0.9442	0.31	0.8702	0.0011
	200	0.9422	0.31	0.8125	0.0016
	300	0.9404	0.31	0.8483	0.0019
	500	0.9389	0.31	0.8214	0.0023
M2M100$_{(400M)}$ / ja → en	10	0.8013	0.17	0.2995	0.1606
	25	0.7353	0.17	0.2994	0.1438
	50	0.7540	0.17	0.3023	0.1603
	75	0.7368	0.16	0.3019	0.1603
	100	0.7453	0.15	0.3072	0.1529
	200	0.7295	0.14	0.2938	0.1787
	300	0.7192	0.13	0.2948	0.1788
	500	0.7110	0.13	0.2756	0.1867
OPT$_{(350M)}$ / OpenWebText	10	0.9438	0.35	0.8824	0.0019
	25	0.9522	0.33	0.8333	2.06×10^{-5}
	50	0.9442	0.27	0	1.86×10^{-5}
	75	0.9477	0.27	0.8	1.03×10^{-5}
	100	0.9460	0.26	0.8	1.86×10^{-5}
	200	0.9487	0.28	0.8571	6.20×10^{-5}
	300	0.9500	0.28	0.8181	1.86×10^{-5}
	500	0.9508	0.29	0.8181	1.86×10^{-5}

being the value we use in our main experiments). We make the following, interesting observations: Coverage on the MT task seems to decrease with an increase in the neighborhood size as prediction set widths get smaller on average, with a neighborhood size around 100 striking a balance between coverage, width, computational cost and SCC / ECG. For LM, coverage seems to be mostly constant, with prediction set width hitting an inflection point for 100 neighbors. We speculate that initially there might be a benefit to considering more neighbors to calibrate $\hat{q}$, but that considering too large neighborhoods might introduce extra noise. While we found 100 to be a solid choice for the purpose of our experiments, we leave more principled ways to determine the neighborhood size to future work.

A.4 Algorithm

We show the algorithm that was schematically depicted in Figure 1 in pseudo-code in Algorithm 1. It mostly requires that we have pre-generated a datastore of latent representations of the model on a held-out set along with their non-conformity scores (in our case, using the score defined in 6 and the FAISS (Johnson et al., 2019) as the datastore architecture). Furthermore, we need to have determined an appropriate value for the temperature τ in advance (see Appendix A.1). Then, the algorithm involves the following steps:

1. Extract the latent encoding for the current time step $\mathbf{z}_t$ from the model. Even though different

Algorithm 1 Non-exchangeable Conformal Language Generation with Nearest Neighbors

Require: Sequence $\mathbf{x}^{(i)}$, model $f_{\boldsymbol{\theta}}$, datastore $\text{DS}(\cdot)$ with model activations collected from held-out set, temperature τ

while generating **do**

▷ 1. Extract latent encoding for current input

$\mathbf{z}_t^{(i)} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x}_t)$

▷ 2. Retrieve K neighbors & non-conformity scores

$\{(\mathbf{z}_1, s_1), \ldots (\mathbf{z}_K, s_K)\} \leftarrow \text{DS}(\mathbf{z}_t)$

▷ 3. Compute weights w_k and normalize

$w_k \leftarrow \exp(-|| \mathbf{z}_t^* - \mathbf{z}_k||_2^2 \ / \ \tau)$

$\tilde{w}_k \ \leftarrow \ w_k / (1 + \sum_{k=1}^{K} w_k)$

▷ 4. Find quantile $\hat{q}$

$\hat{q} \leftarrow \inf\{q \mid \sum_{i=1}^{N} \tilde{w}_i \mathbf{1}\{s_i \leq q\} \geq 1 - \alpha\}$

▷ 5. Create prediction set

$\hat{c} \leftarrow \sup\{c' | \sum_{j=1}^{c'} p_{\boldsymbol{\theta}}(y = \pi(j) | \mathbf{x}^*) < \hat{q}\} + 1$

$\mathcal{C}(\mathbf{x}^*) \leftarrow \{\pi(1), \ldots, \pi(\hat{c})\}$

▷ 6. Generate next token

$y_t \ \leftarrow \ \text{generate}(\mathcal{C}(\mathbf{x}^*))$

end while

options are imaginable, we utilize the activations of the uppermost layer.

2. Retrieve K neighbors and their corresponding non-conformity scores from the datastore.

3. Compute the weights w_k based on the squared l_2 distance between $\mathbf{z}_t$ and its neighbors in the datastore and normalize the weights to obtain $\tilde{w}_k$.

4. Use Equation (3) to find the quantile $\hat{q}$.

5. Use $\hat{q}$ to create prediction sets, for instance the adaptive prediction sets defined in Equation (7).

6. Finally, generate the new token y_t by sampling from the prediction set.

The main computational bottleneck of this algorithm is the retrieval process that fetches the closest neighbors from the datastore during every generation step. However, while not explored further in this work, there are some potential avenues to reduce this load: On the one hand, works such as He et al. (2021b); Martins et al. (2022) have demonstrated ways to reduce the computational load of k-NN based approaches. On other hand, we treat the number of neighbors K fixed during every generation step. However, it seems intuitive that the number of neighbors necessary to create good prediction sets would not be the same for all tokens. Future research could explore setting K dynamically during every time step, thus reducing the overall slowdown.

A.5 Environmental Impact

We track the carbon emissions produced by this work using the `codecarbon` tracking tool (Schmidt et al., 2021; Lacoste et al., 2019; Lottick et al., 2019). The carbon efficiency was estimated to be 0.12 kgCO_2eq / kWh. 159.5 hours of computation were performed on a NVIDIA RTX A6000. Total emissions are estimated to be 6.99 kgCo2eq. All of these values are upper bound including debugging as well as failed or redundant runs, and thus any replication of results will likely be shorter and incur fewer carbon emissions.